Theory of representation learning in cortical neural networks

THÈSE N^{O} 6948 (2016)

PRÉSENTÉE LE 19 FÉVRIER 2016

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
LABORATOIRE DE CALCUL NEUROMIMÉTIQUE (IC/SV)
PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Carlos STEIN NAVES DE BRITO

acceptée sur proposition du jury:

Prof. M. C. Gastpar, président du jury
Prof. W. Gerstner, directeur de thèse
Prof. M. Bethge, rapporteur
Prof. P. Latham, rapporteur
Prof. C. Petersen, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2016

People say: between two opposed opinions the truth lies in the middle.
Not at all! Between them lies the problem, what is unseeable,
eternally active life, contemplated in repose.
—- Goethe

Para Suzana.

Acknowledgements

First and foremost I wish to thank my advisor, Prof. Wulfram Gerstner, for the freedom, the trust and the support he offered me during my entire time at the LCN. Also for his sensible choice of pupils, as my laboratory colleagues were a vibrant and brilliant group, creating an environment rich in ideas, skill and folly. I warmly thank Johanni Brea, Dane Corneil, Moritz Deger, Felipe Gerhard, Laureline Lagiaco, Vasiliki Liakoni, Skander Mensi, Richard Naud, Katy Pegg, Alex Seeholzer, Eszter Vertes, and Lorric Ziegler for the daily fun. A special thanks to Christian Pozzorini, partner in science and in crime.

I thank my gorgeous friends for this eventful swiss era, with a memorable trail of homes, that transformed my self into what it is now: Anaïs, Burçin, Cécile, Dafne, Emile, Emma, Fanny, Flora, Julia, Kula, Leonie, Luciana, Lucile, Magali, Mathilde, Michael, Raph, Sophie, Suzanne, and Tristan. And my friends from home, that never stayed too far away: Ana, Caio, Cazé, Gabriela, Gleyce, Ricardo, Olavo, Testa, Thaise, Thiago, Tom, Vaps, Veruska, and Zeke. I'm particularly grateful to my lifelines, Gustav and Elsa.

And the reason I've done all this, my family. Sofia, Adriano, Elissa and the Stein's, Naves' and Brito's. I've been told that when asked as kid what I wanted to be when I grew up, I answered "a PhD". Well here I am, following our tradition of overexamining everything and leading a fulfilling life because of it. I owe it all to you.

Lausanne, 17th February 2016 C. S. N. B.

Abstract

Our brain continuously self-organizes to construct and maintain an internal representation of the world based on the information arriving through sensory stimuli. Remarkably, cortical areas related to different sensory modalities appear to share the same functional unit, the neuron, and develop through the same learning mechanism, synaptic plasticity. It motivates the conjecture of a unifying theory to explain cortical representational learning across sensory modalities. In this thesis we present theories and computational models of learning and optimization in neural networks, postulating functional properties of synaptic plasticity that support the apparent universal learning capacity of cortical networks.

In the past decades, a variety of theories and models have been proposed to describe receptive field formation in sensory areas. They include normative models such as sparse coding, and bottom-up models such as spike-timing dependent plasticity. We bring together candidate explanations by demonstrating that in fact a single principle is sufficient to explain receptive field development. First, we show that many representative models of sensory development are in fact implementing variations of a common principle: nonlinear Hebbian learning. Second, we reveal that nonlinear Hebbian learning is sufficient for receptive field formation through sensory inputs. A surprising result is that our findings are independent of specific details, and allow for robust predictions of the learned receptive fields. Thus nonlinear Hebbian learning and natural statistics can account for many aspects of receptive field formation across models and sensory modalities.

The Hebbian learning theory substantiates that synaptic plasticity can be interpreted as an optimization procedure, implementing stochastic gradient descent. In stochastic gradient descent inputs arrive sequentially, as in sensory streams. However, individual data samples have very little information about the correct learning signal, and it becomes a fundamental problem to know how many samples are required for reliable synaptic changes. Through estimation theory, we develop a novel adaptive learning rate model, that adapts the magnitude of synaptic changes based on the statistics of the learning signal, enabling an optimal use of data samples. Our model has a simple implementation and demonstrates improved learning speed, making this a promising candidate for large artificial neural network applications. The model also makes predictions on how cortical plasticity may modulate synaptic plasticity for optimal learning.

The optimal sampling size for reliable learning allows us to estimate optimal learning times for a given model. We apply this theory to derive analytical bounds on times for the optimization of synaptic connections. First, we show this optimization problem to have exponentially many saddle-nodes, which lead to small gradients and slow learning. Second, we show that the number of input synapses to a neuron modulates the magnitude of the initial gradient, determining the duration of learning. Our final result reveals that the learning duration increases supra-linearly with the number of synapses, suggesting an effective limit on synaptic connections and receptive field sizes in developing neural networks.

Keywords: representation learning, neural networks, receptive field development, synaptic plasticity, nonlinear Hebbian learning, probabilistic models, sparse coding, independent component analysis, spiking neurons, stochastic optimization, adaptive learning rates, learning dynamics.

Résumé

Notre cerveau s'auto-organise constamment pour construire et entretenir une représentation interne du monde en se basant sur l'information arrivant des stimuli sensoriels. De manière remarquable, des régions corticales liées à différentes modalités sensorielles semblent partager la même unité fonctionnelle de base, le neurone, et se développer via le même mécanisme d'apprentissage, la plasticité synaptique. Ceci motive la conjecture d'une théorie unique pour expliquer l'apprentissage des représentations corticales de toutes les modalités sensorielles. Nous présentons dans cette thèse plusieurs théories et modèles computationnels d'apprentissage et d'optimisation dans les réseaux neuronaux, qui postulent certaines propriétés de la plasticité synaptique, et soutiennent l'existence d'une capacité universelle d'apprentissage dans les réseaux corticaux.

Dans les dernières décennies, une grande variété de théories et de modèles ont été proposés pour décrire la formation des champs récepteurs dans les zones sensorielles. Cela inclue à la fois des modèles normatifs (Sparse Coding) et inductifs, où, par exemple, la plasticité dépend des instants d'occurrence des potentiels d'action. Dans cette thèse, différentes tentatives d'explications sont réconciliées en démontrant qu'un seul et unique principe est suffisant pour expliquer le développement des champs récepteurs. D'abord, nous démontrons que beaucoup de modèles du développement sensoriel ne font qu'implémenter des variations du même principe de base : la règle de Hebb non-linéaire. Ensuite, nous prouvons que la théorie de Hebb non-linéaire est suffisante pour expliquer la formation des champs récepteurs à partir des données sensorielles. Un aspect surprenant de nos résultats est leur robustesse vis-à-vis des spécificités individuelles des modèles, ce qui permet de prédire de manière fiable les champs récepteurs acquis. Ainsi, la théorie de Hebb non-linéaire, combinée aux propriétés statistiques des stimuli naturels, peut expliquer de nombreux aspects de la formation des champs récepteurs, et ce, pour de multiples modèles et modalités sensorielles.

La théorie de Hebb tend à prouver que la plasticité synaptique peut être interprétée comme une procédure d'optimisation, implémentant une descente de gradient stochastique. En effet, lors d'une descente de gradient stochastique, de nouvelles données arrivent successivement, comme des flux sensoriels. Cependant, les échantillons de données ne contiennent individuellement que très peu d'information permettant de déterminer le bon signal d'apprentissage, et estimer la quantité d'échantillons nécessaire à l'obtention d'ajustements synaptiques fiables est un problème fondamental. Grâce à la théorie de l'estimation, nous développons un nouveau modèle à taux d'apprentissage adaptatif, qui ajuste l'amplitude des changements synaptiques en se basant sur les statistiques du signal d'apprentissage, permettant une uti-

lisation optimale des données. Notre modèle est simple à implémenter et présentent des temps d'apprentissages plus courts que les modèles classiques, ce qui en fait un candidat prometteur pour des applications à de larges réseaux de neurones artificiels. De plus, des prédictions concernant la modulation de la plasticité synaptique par la plasticité corticale pour un apprentissage optimal peuvent en être dérivées.
L'estimation de la durée minimale de l'échantillonnage permettant un apprentissage fiable nous donne la possibilité d'estimer les temps d'apprentissage optimaux pour un modèle donné. Nous appliquons ici cette théorie pour dériver analytiquement des majorants du temps d'optimisation des connexions synaptiques. Pour commencer, nous montrons que ce problème d'optimisation présente un nombre de points-cols qui croît exponentiellement avec le nombre de neurones, menant à des gradients faibles et à un apprentissage lent. Dans un deuxième temps, nous montrons que le nombre de synapses d'entrée d'un neurone module la magnitude du gradient initial, fixant la durée de l'apprentissage. Enfin, le résultat final de cette thèse révèle que la durée de l'apprentissage augmente supra-linéairement avec le nombre de synapses, suggérant des limites effectives au nombre de connexions synaptiques et à la taille des champs récepteurs dans des réseaux neuronaux en phase de développement.

Mots clefs : apprentissage de représentations, réseaux neuronaux, développement de champs récepteurs, plasticité synaptique, théorie de Hebb non-linéaire, modèles probabilistes, codage creux, analyse en composantes indépendantes, réseaux de neurones à impulsions, optimisation stochastique, taux d'apprentissage adaptatif, dynamique de l'apprentissage.

Contents

List of Figures

List of Tables

1 Introduction

Our knowledge about the external world is delivered by sensory organs, however our subjective perception does not reside in the eyes or ears, but is a construction made possible by a representation of the data in the cortex (James, 1890; Merleau-Ponty, 1945). By inspecting how different brain areas respond to external stimuli, we may correlate them with specific cognitive abilities, such as object recognition (Grill-Spector et al., 1998; Quiroga et al., 2005; Desimone, 1991), color perception (Conway and Tsao, 2006; Komatsu et al., 1992), language understanding (Mazoyer et al., 1993; Price, 2010) or spatial awareness (Andersen et al., 1985; Duhamel et al., 1997).

This diversity of cognitive functions is in striking contrast to the apparent homogeneity in the neural substrate across cortical areas (Mountcastle, 1978). All regions display neurons as functional units, arranged in stereotypical layers, and the essential difference between them is to which properties of the external world they respond to. This difference in receptive field properties can be explained by differentiated connections. And, in turn, the receptive field of a neuron further along the processing stream will be shaped by the receptive field properties of the neurons that project to it.

Sensory cortices require sensory input in early life to develop (Wiesel and Hubel, 1963b; Crair et al., 1998), suggesting that our representation of the external world is not genetically predetermined but is learned through experience (Viskontas et al., 2009). And yet, plasticity rules that govern the self-organization of synaptic connections are also remarkably consistent across sensory areas (Caporale and Dan, 2008). These findings suggest the existence of learning principles common across the cortex, flexible enough to adapt synaptic connections to represent very different input modalities.

Most proposals of universal learning algorithms come from the field of artificial neural networks. In the early days of artificial intelligence, the perceptron model showed that a plasticity rule could make a network of simplified neurons learn to discriminate between input categories (Rosenblatt, 1958). Multi-layer versions of these networks were proven to have universal representation capabilities and later an algorithm was developed that could train these for a variety of categorization tasks (Rumelhart et al., 1988; LeCun and Bengio, 1995). In recent decades probabilistic models were also developed, in which stochastic neural networks learn to represent the statistical distribution of the input, generalizing their properties to generate new data samples that haven't been seen before (Hinton et al., 2006). These networks also have universal properties, being able to represent any probability distribution, given enough neurons (Le Roux and Bengio, 2008).

However impressive, these models are only loosely inspired by biological networks, and many of their properties are not compatible with facts known about cortical neurons, synaptic plasticity and network dynamics. Thus, in parallel to these functional networks, biological models of plasticity were developed taking into account the evidence and constraints coming from experimental neuroscience.

The ideas of psychologist Donald Hebb were a precursor for biological modeling of learning (Hebb, 1952). He proposed that associative memories, in which co-occurring concepts become linked, could be implemented by reinforcing the connections between neurons that represented each concept. This became known as the principle of Hebbian plasticity: neurons that fire together, wire together. Plasticity models based on the Hebbian principle thrived, with increasing support from both theory and experiments (Oja, 1989; Markram et al., 1997; Pfister and Gerstner, 2006; Cooper and Bear, 2012).

This thesis builds on these previous efforts, presenting new theoretical results on synaptic plasticity rules. By considering simplified models, we give support to cortical plasticity as a universal learning mechanism, designed for the search of statistical structure.

1.1 Sensory representation

1.1.1 Functional specialization

Distinct brain regions are involved in specific cognitive processes. The study of cortical functional specialization has its origins in the infamous theory of phrenology by Franz Gall, in which personality traits were related to bumps in specific regions of the skull (Kandel et al., 2000). Apart from these early pseudo-scientific attempts, solid discoveries were possible by observing correlations between cognitive impairment and brain lesions.

Paul Broca discovered that patients with lesions in a region of the frontal lobe (now Broca's area) present deficit in language production, while Carl Wernicke found that another brain area (now Wernicke's area) was necessary for language comprehension. Lesion studies continue to this day to expose the modularity of cortical functioning. As beautifully portrayed by Oliver Sacks (1998), patients can display curiously specific cognitive impairments, such as face blindness, or ability to write but not to read.

The study of functional specialization was revolutionized in the 1990's by the development of functional magnetic resonance imaging (fMRI). With fMRI, brain activity is measured indirectly from local changes in blood oxygenation levels. By measuring these changes during particular tasks, one can find the neural correlates for any cognitive function imaginable (Friston et al., 1998; Huettel et al., 2004).

A common criticism of such neural correlates is that they only map function to brain area, as a novel form of phrenology, but do not give insights on how the underlying neural circuits can produce such functions (Friston, 2002).

1.1.2 Single neuron receptive fields

The study of single neuron activations complement these findings by characterizing the stimuli that trigger a neuronal response. In these experiments, electric wires are introduced in the cortical area of interest to measure the electric activity of nearby neurons during sensory stimuli. The sensory stimuli that triggers the neuron define its receptive field. Hubel and Wiesel (1959) provided the first of such characterizations for cortical neurons, showing that neurons in the primary visual cortex (V1) in the cat were selectively activated by oriented bars at specific locations and angles of the visual field, named simple cells. It was a revealing discovery, showing that neurons represented highly specific patterns of the external world and that, at least in these cases, the patterns were intelligible. More discoveries followed, and neurons with more complex and abstract receptive fields were found, including neurons in the monkey's temporal cortex selective to images of faces (Rolls et al., 1992) and hands (Gross, 2008), or cells in the rats' hyppocampus responsive to specific spatial coordinates (O'Keefe, 1979). And these receptive fields were not static, being continuously shaped by experience (Wiesel and Hubel, 1963a; Foster and Wilson, 2006), with particularly plastic periods occurring in early life in the so called critical period of a sensory area (Hensch, 2005).

As an invasive technique, single neuron characterization is restricted to animal studies. Nevertheless, in rare occasions, patients that will undergo brain surgery volunteer for short experiments, allowing for receptive field characterizations in the human brain. In one of these experiments, patients were shown a collections of images, including landmarks, animals and television celebrities, while neurons from the medial temporal lobe were recorded (Quiroga et al., 2005). Strikingly, those neurons revealed receptive fields with an unprecedented degree of abstraction. One neuron was selective to images of the actress Jennifer Aniston, but not to any of the other actresses. On the other hand, it was responsive to all of her pictures, in different poses and clothing, and also to her spoken or written name. Other neurons were found selective to Sadam Hussein, the Sidney Opera house and Pythagora's theorem. These experiments also showed that neural representations are highly plastic, finding neurons with selective response to the experimenter known to the patient only for a few days (Viskontas et al., 2009).

The examples we have listed are a biased selection of neurons with receptive fields that are easy to interpret. Primary sensory neurons may be summarized as linear filter of sensory inputs, while high level neurons may be described by abstract concepts. However these are only first approximations, and receptive field descriptions have many caveats. First, the modularity implied by the functional specialization of brain areas is rather nebulous, and is not as simple as presented this far. For instance, visual processing can be modulated by auditory stimuli (Frassinetti et al., 2002). Second, under closer inspection even V1 receptive fields have complex properties (Olshausen and Field, 2005). Third, abstract concepts also have non-trivial selectivity properties, for instance the Jennifer Aniston neuron did not respond to images in which Brad Pitt was also present (Quiroga et al., 2005). Forth, the majority of neurons are in intermediate levels of abstraction, and have receptive fields that may be hard to characterize even at an approximate level (Nandy et al., 2013). These objections suggest caution in the description of receptive fields, but are consistent with the notion that neurons respond to complex patterns and that investigating individual receptive fields can give insight about representational properties of sensory networks.

1.1.3 A modern neuron doctrine

Horace Barlow took receptive field findings as basis for the formulation of a neuron doctrine, on how to understand sensory representations based on single neurons (Barlow, 1972). His proposal shows incredible foresight, and we transcribe it here:

1. To understand nervous function one needs to look at interactions at a cellular level, rather than either a more macroscopic or microscopic level, because behaviour depends upon the organized pattern of these intercellular interactions.

2. The sensory system is organized to achieve as complete a representation of the sensory stimulus as possible with the minimum number of active neurons.

3. Trigger features of sensory neurons are matched to redundant patterns of stimulation by experience as well as by developmental processes.

4. Perception corresponds to the activity of a small selection from the very numerous high-level neurons, each of which corresponds to a pattern of external events of the order of complexity of the events symbolized by a word.

5. High impulse frequency in such neurons corresponds to high certainty that the trigger feature is present.

Although single neurons are certainly not the only level of abstraction relevant for understanding sensory systems, his doctrine is remarkably consistent with the assumptions made by models studied in this thesis. Consistent with his first statement, we consider phenomenological models of neurons and synapses, with an agnostic position as to how these are implemented at a molecular level (Gerstner et al., 2014). His second statement underlies sparse coding models (Olshausen and Field, 1996), lateral inhibition (Vogels et al., 2011) and neural adaptation

(Pozzorini et al., 2013), with sensory networks representing its input with maximum efficiency. His third statement is the central assumption of the Hebbian models we investigate, in which neurons learn to represent patterns that occur often in the input during sensory development (Oja, 1989; Cooper and Bear, 2012). His forth hypothesis clearly predicts the Jennifer Aniston neurons and the like (Quiroga et al., 2005), and also the highly abstract neurons seen in artificial neural networks (Le, 2013). Finally, his last statement is consistent with how neural activity is interpreted in some probabilistic models of neural activity, such as sparse coding models (Olshausen and Field, 1996) and Boltzmann machines (Hinton, 2002; Buesing et al., 2011).

1.2 Sensory development

1.2.1 Receptive fields and natural input statistics

As sensory experience shaped sensory representations, novel results on sensory experience drove the investigation of what properties of the sensory data were relevant in receptive field development. A breakthrough came from the studies of Field (1994), which revealed that receptive fields of simple cells could be related to the statistics of natural images. Considering a set of image patches $\{\mathbf{x}_i\}$, and synaptic projections $\mathbf{w}$, the input to the projected neuron for each image patch is given by $I = \mathbf{w}^T \mathbf{x}_i$. He analyzed the statistical distribution of I for different choices of patterns $\mathbf{w}$, and showed that kurtosis (the forth moment of the distribution, $\langle I^4 \rangle$), was higher for oriented edges. Since simple cells were selective to oriented edges, it suggested that neurons may be driven to adapt their synapses to projections that maximize higher-order statistics. It was evidence for a statistical formalism that could replace the heuristic argument that sensory neurons learned to represent patterns that occurred often.

1.2.2 Sparse coding and probabilistic models

This statistical insight gave rise to a normative theory, known as sparse coding (Olshausen and Field, 1996). The objective of sensory neurons would be that of encoding a large statistical ensemble of inputs with minimal information loss, while using minimal neural activity. Since projections of the input distribution with higher kurtosis have sparser activity, it was expected that simple cells could develop from a model that enforced these assumptions. The sparse coding model is formalized with latent neurons $\mathbf{y}$, which represent the input $\mathbf{x}$, constructed as a linear sum, $x_i = \mathbf{w}_i \mathbf{y}$. Learning is then defined as an optimization problem, the search for synaptic weights $\mathbf{w}$ that minimize the reconstruction error of the input on average over many inputs, while penalizing for neural activity,

$$E = \frac{1}{2} ||\mathbf{x} - \mathbf{W}\mathbf{y}||^2 + \lambda \sum S(y_k) \tag{1.1}$$

where $S(.)$ is the sparsity penalty function. Although not biologically plausible (e.g. the learning rule violates locality), the model is neural-like, composed of two layers, an input layer $\mathbf{x}$ and a processing layer **y**, connected through a matrix of synaptic connections **W** (see Chapter 2).

The same formulation can be interpreted as a generative probabilistic model, a directed graphical model in which the input **x** is generated from latent variables **y**, with a generative distribution given by

$$P(\mathbf{x}|\mathbf{y},\mathbf{W}) = \frac{1}{Z_x}\, e^{-\frac{||\mathbf{x}-\mathbf{W}^T\mathbf{y}||^2}{2\sigma^2}} \tag{1.2}$$

and latent variables have a prior distribution given by

$$P(\mathbf{y}) = \frac{1}{Z_y} \prod_j e^{-\alpha S(y_j)} \tag{1.3}$$

where α and σ are constants, and Z_x and Z_y are normalization factors, known as partition functions.

In this interpretation (Olshausen and Field, 1997), the learning process performs maximum likelihood estimation, maximizing the probability that the input dataset has been generated by the model,

$$\mathbf{w}^* = \text{argmax}_{\mathbf{w},\ |w|_2=1}\ \ell = logP(\mathbf{x},\mathbf{y}|\mathbf{W}) \propto -\frac{1}{2}||\mathbf{x}-\mathbf{W}^T\mathbf{y}||^2 - 2\alpha\sigma^2\sum_j S(y_j) + c. \tag{1.4}$$

When the sparse coding model was optimized for natural image patches as input, oriented edges developed as receptive fields **w** (Olshausen and Field, 1996). It was an exciting result, finding simultaneous connections between receptive fields and natural statistics, between neural activity and probabilistic representations, and between learning and statistical optimization.

Since then further experimental evidence have supported the sparse coding principle. Neural activity in the cortex has been observed to be very sparse (Barth and Poulet, 2012), and sensory development appears to make spontaneous neural activity more similar to activity driven by natural inputs (Berkes et al., 2011), proposing sparse coding as a general principle in sensory development (Olshausen and Field, 2004).

1.2.3 Hebbian plasticity and optimization

Normative approaches such as sparse coding derive their model neurons and learning rules from objective functions, and in most cases the resulting model does not fit with known experimental findings. For instance, the sparse coding model has learning mechanisms that are non-local, in the sense that synapses from one neuron must have information about synapses from other neurons to estimate its update (see Chapter 2).

We may instead follow a bottom up approach and study simple, yet biologically plausible, local plasticity rules. From Hebb's principle of learning by association, we may define a plasticity rule in which synapses w_{ji} are potentiated when the pre- and post-synaptic neurons are co-active, $\Delta w_{ji} \propto x_i \, y_j$. However, this learning rule has an obvious handicap. If neural activations are non-negative, then weights only potentiate, and will diverge to infinity. One way to mitigate this problem is to include a depressive term, which can counter-balance potentiation and lead to a stable equilibrium.

In Oja's learning rule (Oja, 1989), the depression term depends quadratically on the activity of the post-synaptic neuron

$$\Delta w_{ji} \propto x_i \, y - x_i \, y^2 \tag{1.5}$$

and the post-synaptic neuron is modeled with a linear model, $y = \mathbf{w}^T \mathbf{x}$. As such, if the synaptic weights start to increase, the post-synaptic activity increases, which in turn leads to synaptic depression. Interestingly, Oja showed that this learning rule was an implementation of principal component analysis, proving that synapses converge to the input projection with highest variance,

$$w^* = \text{argmax}_{\mathbf{w},\ |\mathbf{w}|_2=1} \left\langle (\mathbf{w}^T \mathbf{x})^2 \right\rangle \tag{1.6}$$

This illustrates how a very simple Hebbian learning rule can implement well defined statistical optimization algorithms.

A second prominent model is the BCM model. This is a nonlinear Hebbian model, also designed to be stable (Bienenstock et al., 1982b):

$$\Delta w_{ji} \propto x_i \, y_j^2 - \theta \, x_i \, y_j \tag{1.7}$$

where θ is a dynamic variable, $\theta = \left\langle y_i^2 \right\rangle$, which guarantees that synaptic weights do not diverge. This model is also implementing an optimization function (Intrator and Cooper, 1992), but in this case it depends on higher order statistics,

$$w^* = \text{argmax}_{\mathbf{w}} \left\langle y^3 \right\rangle - c \left\langle y^2 \right\rangle^2 \tag{1.8}$$

relating it to objective functions based on kurtosis. Like the sparse coding model, it develops simple-cell like receptive fields for natural image inputs (Law and Cooper, 1994) . The BCM model has gathered increasing support as experiments have observed its quadratic dependency on the post-synaptic activity (Cooper and Bear, 2012; Pfister and Gerstner, 2006). Thus the BCM model exhibits simultaneously higher-order statistical properties, stable behavior and adequacy to plasticity experiments.

Oja's and BCM models are instances of a broad class of Hebbian models that implement a statistical optimization procedure known as projection pursuit (Friedman, 1987; Intrator and Cooper, 1992; Fyfe and Baddeley, 1995). In projection pursuit, one searches for synaptic weights that maximize a given objective function $F(.)$,

$$w^* = \mathrm{argmax}_{\mathbf{w}} \left\langle F(\mathbf{w}^T\mathbf{x}) \right\rangle \tag{1.9}$$

We consider then gradient ascent as the optimization procedure, in which parameters (synaptic weights in this case) are updated in small steps in the direction that increases the objective function (the gradient),

$$\Delta\mathbf{w} \propto \frac{\partial \left\langle F(\mathbf{w}^T\mathbf{x}) \right\rangle}{\partial \mathbf{w}} \propto \left\langle \mathbf{x}\, f(\mathbf{w}^T\mathbf{x}) \right\rangle \tag{1.10}$$

where $F = \int f$. Since the expectation over the whole input dataset is not available, an online optimization procedure is considered instead. In stochastic gradient ascent, parameters are updated based on gradients obtained from single input samples,

$$\Delta\mathbf{w} \propto \mathbf{x}\, f(\mathbf{w}^T\mathbf{x}) \tag{1.11}$$

which for small enough updates will on average approximate its offline gradient ascent equivalent. This general formulation of learning rules defines the class of nonlinear Hebbian learning rules that implement projection pursuit (Fyfe and Baddeley, 1995; Hyvarinen and Oja, 1998).

1.2.4 Spiking models

The models presented so far ignore a central aspect of neural activity: time. They formalize neural activity in discrete time steps, and neuron output activity as analog values. Instead, cortical neurons are better described as continuous-time dynamical systems, which integrate time-varying currents and have all-or-none output signals, known as spikes or action potentials.

We consider the leaky integrate-and-fire as our standard dynamical systems model for spiking neurons. The integration equation is given by

$$\tau_m \frac{\partial u}{\partial t} = -u(t) + RI(t) \tag{1.12}$$

where $I(t)$ is the time-varying input current and $u(t)$ is time-varying membrane potential. The resistance R and the membrane time-scale τ_m are properties of the neuron. The neuron fires, or spikes, at time t_f if the membrane potential crosses a threshold θ, $v(t_f) = \theta$, and immediately thereafter the membrane potential is reset to a value u_R. There are many possible variations of spiking models, including for instance adaptation currents or stochastic firing (Gerstner et al., 2014).

Although spiking neuron models can implement complex dynamics (Naud et al., 2008), we may instead simplify the description of the neural activity by its average statistics. By defining the firing rate y_t as the number of spikes between time t and $t + \Delta t$, and the input current as the average current in the same time interval, we can approximate the spiking model by a rate model. This simplification is possible for any spiking model variation. Whenever the average statistics are the important variables, it allows us to generalize results obtained from rate neurons to spiking models.

A final, but highly important aspect of biological synaptic plasticity, is that cortical plasticity is dependent on the timing of spikes. In spike-timing dependent plasticity model (STDP), synaptic potentiation occurs when pre-synaptic spikes precede post-synaptic spikes, while depression occurs when post-synaptic spikes come first (Gerstner et al., 1996; Markram et al., 1997). For our purposes however, we may again perform a reduction to a rate model, by assuming that spikes are generated by a Poisson process, so that on average only the rate of the two neurons, measured over a given time scale, will influence learning (Pfister and Gerstner, 2006). In the case of the standard STDP model, dependent only on pairs of pre- and post-synaptic spikes, this reduces the most simple spike-timing sensitive rules to a linear Hebbian model, $\Delta w_{ji} \propto x_i \, y_j$.

Plasticity experiments also inspected more complex combinations of pre- and post-synaptic spikes (Froemke and Dan, 2002). These revealed the presence of nonlinear effects, in which plasticity may depend on higher-order combinations of spikes, such as triplets of spikes. Phenomenological modeling of such data revealed that minimal models of STDP depend on one pre-synaptic spike, but are modulated by the timing of two post-synpatic spikes (Pfister and Gerstner, 2006). Remarkably, this implies that, under the Poisson assumption, this model is equivalent to the BCM rate model, without the dynamic threshold (Pfister and Gerstner, 2006). This opens the way to assign to spiking plasticity models all the functional properties studied in their rate equivalents, including receptive field development (Clopath et al., 2010).

1.2.5 Unifying normative and biological models

While top-down normative approaches, such as sparse coding, have a strong appeal by providing a conceptual framework to understanding representation and learning, they do not possess the biological plausibility of bottom-up biological models, such as spiking networks and STDP. Nevertheless, there are commonalities between both types of models that invite an attempt at unification. These models are neural-like, with Hebbian-like learning rules, and are able to learn similar receptive fields from natural inputs. We start this thesis by demonstrating that these impressions can be formalized, revealing nonlinear Hebbian learning as the common principle behind these different approaches.

2 Nonlinear Hebbian Learning as a Unifying Principle in Receptive Field Formation

The development of sensory receptive fields has been modeled in the past by a variety of models including normative models such as sparse coding or independent component analysis and bottom-up models such as spike-timing dependent plasticity or the Bienenstock-Cooper-Munro model of synaptic plasticity. Here we show that the above variety of approaches can all be unified into a single common principle, namely nonlinear Hebbian learning. When nonlinear Hebbian learning is applied to natural images, receptive field shapes were strongly constrained by the input statistics and preprocessing, but exhibited only modest variation across different choices of nonlinearities in neuron models or synaptic plasticity rules. Neither overcompleteness nor sparse network activity are necessary for the development of localized receptive fields. The analysis of alternative sensory modalities such as auditory models or V2 development lead to the same conclusions. In all examples, receptive fields can be predicted a priori by reformulating an abstract model as nonlinear Hebbian learning. Thus nonlinear Hebbian learning and natural statistics can account for many aspects of receptive field formation across models and sensory modalities.

2.1 Introduction

Neurons in sensory areas of the cortex are optimally driven by stimuli with characteristic features that define the ’receptive field’ of the cell. While receptive fields of simple cells in primary visual cortex (V1) are localized in visual space and sensitive to the orientation of light contrast (Hubel and Wiesel, 1959), those of auditory neurons are sensitive to specific time-frequency patterns in sounds (Miller et al., 2002). The concept of a receptive field is also useful when studying higher-order sensory areas, for instance when analyzing the degree of selectivity and invariance of neurons to stimulus properties (DiCarlo et al., 2012; Freeman and Simoncelli, 2011).

The characteristic receptive fields of simple cells in V1 have been related to statistical properties of natural images (Field, 1994). These findings inspired various models, based on principles as diverse as sparse sensory representations (Olshausen and Field, 1996), optimal information transmission (Bell and Sejnowski, 1997), or synaptic plasticity (Law and Cooper, 1994). Several studies highlighted possible connections between biological and normative justifications of sensory receptive fields (Rehn and Sommer, 2007; Clopath et al., 2010; Savin et al., 2010; Zylberberg et al., 2011), not only in V1, but also in other sensory areas (Olshausen and Field, 2004), such as auditory (Smith and Lewicki, 2006; Saxe et al., 2011) and secondary visual cortex (V2) (Lee et al., 2007).

Since disparate models appear to achieve similar results, the question arises whether there exists a general underlying concept in unsupervised learning models (Saxe et al., 2011; Yamins et al., 2014). Here we show that the principle of nonlinear Hebbian learning is sufficient for receptive field development under rather general conditions. The nonlinearity is defined by the neuron's f-I curve combined with the nonlinearity of the plasticity function. The outcome of such nonlinear learning is equivalent to projection pursuit (Friedman, 1987; Oja et al., 1991; Fyfe and Baddeley, 1995), which focuses on features with non-trivial statistical structure, and therefore links receptive field development to optimality principles.

Here we unify and broaden the above concepts and show that plastic neural networks, sparse coding models and independent component analysis can all be reformulated as nonlinear Hebbian learning. For natural images as sensory input, we find that a broad class of nonlinear Hebbian rules lead to orientation selective receptive fields, and explain how seemingly disparate approaches may lead to similar receptive fields. The theory predicts the diversity of receptive field shapes obtained in simulations for several different families of nonlinearities. The robustness to model assumptions also applies to alternative sensory modalities, implying that the statistical properties of the input strongly constrain the type of receptive fields that can be learned. Since the conclusions are robust to specific properties of neurons and plasticity mechanisms, our results support the idea that synaptic plasticity can be interpreted as nonlinear Hebbian learning, implementing a statistical optimization of the neuron's receptive field properties.

2.2 Results

2.2.1 The effective Hebbian nonlinearity

In classic rate models of sensory development (Miller et al., 1989; Law and Cooper, 1994; Olshausen and Field, 1996), a first layer of neurons, representing the sensory input $\mathbf{x}$, is connected to a downstream neuron with activity y, through synaptic connections with weights $\mathbf{w}$ (Fig. 2.1a). The response to a specific input is $y = g(\mathbf{w}^T\mathbf{x})$, where g is the frequency-current (f-I) curve. In most models of Hebbian plasticity (Bienenstock et al., 1982a; Gerstner et al., 2014), synaptic changes $\Delta\mathbf{w}$ of the connection weights depend on pre- and post-synaptic activity, with a linear dependence on the pre-synaptic and a nonlinear dependence on the post-synaptic activity, $\Delta\mathbf{w} \propto \mathbf{x}\, h(y)$, in accordance with models of pairing experiments (Pfister and Gerstner, 2006; Clopath et al., 2010). The learning dynamics arise from a combination of the neuronal f-I curve $y = g(\mathbf{w}^\mathbf{T}\mathbf{x})$ and the Hebbian plasticity function $\Delta\mathbf{w} \propto \mathbf{x}\, h(y)$:

$$\Delta\mathbf{w} \propto \mathbf{x}\, h(g(\mathbf{w}^\mathbf{T}\mathbf{x})) = \mathbf{x}\, f(\mathbf{w}^\mathbf{T}\mathbf{x}) \tag{2.1}$$

where we define the *effective Hebbian nonlinearity* $f := h \circ g$ as the composition of the nonlinearity in the plasticity rule and the neuron's f-I curve. In an experimental setting, the pre-synaptic activity x is determined by the set of sensory stimuli (influenced by, e.g., the rearing conditions during sensory development (Wiesel and Hubel, 1963a)). Therefore, the evolution of synaptic strength, Eq. 2.1, is determined by the effective nonlinearity f and the statistics of the input $\mathbf{x}$.

Many existing models can be formulated in the framework of Eq. 2.1. For instance, in a classic study of simple-cell formation (Law and Cooper, 1994), the Bienenstock-Cooper-Munro (BCM) model (Bienenstock et al., 1982a) has a quadratic plasticity nonlinearity, $h(y) = y(y-\theta)$, with a variable plasticity threshold θ, and a sigmoidal f-I curve, $\sigma(\mathbf{w}^T\mathbf{x})$, which combine into nonlinear Hebbian learning dynamics, $\Delta\mathbf{w} \propto \mathbf{x}\, h(\sigma(\mathbf{w}^T\mathbf{x}))$.

More realistic cortical networks have dynamical properties which are not accounted for by rate models. By analyzing state-of-the-art models of cortical neurons and synaptic plasticity, we inspected whether plastic spiking networks can be reduced to nonlinear Hebbian learning. We considered a generalized leaky integrate-and-fire model (GIF), which includes adaptation, stochastic firing and predicts experimental spikes with high accuracy (Pozzorini et al., 2013), and we approximate its f-I curve by a linear rectifier, $g(u) = a(u-\theta)_+$, with slope a and threshold θ (Fig. 2.1b).

As a phenomenological model of synaptic plasticity grounded on experimental data (Sjostrom et al., 2001), we implemented triplet spike-timing dependent plasticity (STDP) (Pfister and Gerstner, 2006). In this STDP model, the dependence of long-term potentiation (LTP) upon two post-synaptic spikes induces in the corresponding rate model a quadratic dependence on the post-synaptic rate, while long-term depression (LTD) is linear. The resulting rate plasticity (Pfister and Gerstner, 2006) is $h(y) = y^2 - by$, with an LTD factor b (post-synaptic activity threshold separating LTD from LTP, Fig. 2.1c), similar to the classic BCM model (Bienenstock et al., 1982a; Law and Cooper, 1994).

Composing the f-I curve of the GIF with the $h(y)$ for the triplet plasticity model, we have an approximation of the effective learning nonlinearity $f = h \circ g$ in cortical spiking neurons (Fig 2.1d), that can be described as a quadratic rectifier, with LTD threshold given by $\theta_1 = \theta$ and LTP threshold given by $\theta_2 = \theta + b/a$. Interestingly, the f-I slope a and LTD factor b are redundant parameters of the learning dynamics: only their ratio counts in nonlinear Hebbian plasticity. Metaplasticity can control the LTD factor (Pfister and Gerstner, 2006; Turrigiano, 2011), thus regulating the LTP threshold.

If one considers a linear STDP model (Song et al., 2000; Gerstner et al., 1996) instead of the triplet STDP (Pfister and Gerstner, 2006), the plasticity curve is linear (Gerstner et al., 2014), as in standard Hebbian learning, and the effective nonlinearity is shaped by the properties of the f-I curve (Fig. 2.2a).

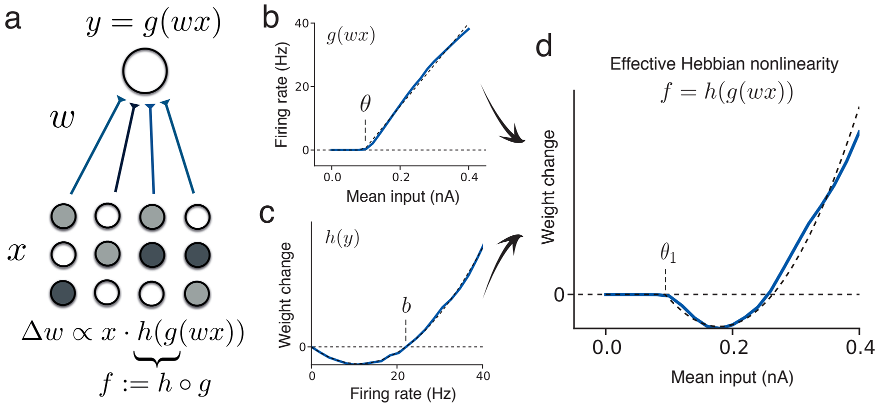

Figure 2.1 – **The effective Hebbian nonlinearity of plastic cortical networks.** (**a**) Receptive field development between an input layer of neurons with activities x_i, connected by synaptic projections w_i to a neuron with firing rate y, given by an f-I curve $y = g(\mathbf{w}^T\mathbf{x})$). Synaptic connections change according to a Hebbian rule $\Delta w_i \propto x_i \ h(y)$. (**b**) f-I curve (blue) of a GIF model (Pozzorini et al., 2013) of a pyramidal neurons in response to step currents of 500 ms duration (dashed line: piece-wise linear fit, with slope $a = 143$ Hz/nA and threshold $\theta = 0.08$ nA). (**c**) Plasticity function of the triplet STDP model (Pfister and Gerstner, 2006) (blue), fitted to visual cortex plasticity data (Sjostrom et al., 2001; Pfister and Gerstner, 2006), showing the weight change Δw_i as a function of the post-synaptic rate y, under a constant pre-synaptic stimulation x_i (dashed line: fit by quadratic function, with LTD factor $b = 22.1$ Hz). (**d**) The combination of the f-I curve and plasticity function generates the effective Hebbian nonlinearity (dashed line: quadratic nonlinearity with LTD threshold $\theta_1 = 0.08$ nA, LTP threshold $\theta_2 = 0.23$ nA).

2.2.2 Sparse coding as nonlinear Hebbian learning

Beyond phenomenological modeling, normative principles that explain receptive fields development have been one of the goals of theoretical neuroscience (Dayan and Abbott, 2001). Sparse coding (Olshausen and Field, 1996) starts from the assumptions that V1 aims at maximizing the sparseness of the activity in the sensory representation, and became a well-known normative model to develop orientation selective receptive fields (Rehn and Sommer, 2007; Zylberberg et al., 2011; Olshausen and Field, 2004). We demonstrate that the algorithm implemented in the sparse coding model is in fact a particular example of nonlinear Hebbian learning.

The sparse coding model aims at minimizing an input reconstruction error $E = \frac{1}{2}||\mathbf{x} - \mathbf{W}\mathbf{y}||^2 +$

$\lambda S(\mathbf{y})$, under a sparsity constraint S with relative importance $\lambda > 0$. For K hidden neurons y_j, such a model implicitly assumes that the vector $\mathbf{w_j}$ of feed-forward weights onto neuron j are mirrored by hypothetical "reconstruction weights", $\mathbf{W} = [\mathbf{w}_1 \ldots \mathbf{w}_K]$. The resulting encoding algorithm can be recast as a neural model (Rozell et al., 2008), if neurons are embedded in a feedforward model with lateral inhibition, $\mathbf{y} = g(\mathbf{w}^T\mathbf{x} - \mathbf{v}^T\mathbf{y})$, where v are inhibitory recurrent synaptic connections (see Methods). In the case of a single output neuron, its firing rate is simply $y = g(\mathbf{w}^T\mathbf{x})$. The nonlinearity g of the f-I curve is threshold-like, and determined by the choice of the sparsity constraint (Rozell et al., 2008), such as the Cauchy, L_0 , or L_1 constraints (Fig 2.2a, see Methods).

If weights are updated through gradient descent so as to minimize E, the resulting plasticity rule is Oja's learning rule (Oja, 1982), $\Delta\mathbf{w} \propto \mathbf{x}\, y - \mathbf{w}\, y^2$. The second term $-\mathbf{w}\, y^2$ has a multiplicative effect on the strength of synapses projecting onto the same neuron (weight rescaling), but does not affect the receptive field shape, whereas the first term $\mathbf{x}\, y$ drives feature selectivity and receptive field formation. Together, these derivations imply that the one-unit sparse coding algorithm can be implemented by an effective nonlinear Hebbian rule combined with weight normalization. Although the plasticity mechanism is linear, $\Delta\mathbf{w} \propto \mathbf{x}\, y$, a nonlinearity arises from the f-I curve, $y = g(\mathbf{w}^T\mathbf{x})$, so that the effective plasticity is

$$\Delta\mathbf{w} \propto \mathbf{x}\, g(\mathbf{w}^T\mathbf{x}) \tag{2.2}$$

This analysis reveals an equivalence between sparse coding models and neural networks with linear plasticity mechanisms, where the sparsity constraint is determined by the f-I curve g.

Similarly, algorithms performing independent component analysis (ICA), a model class closely related to sparse coding, also perform effective nonlinear Hebbian learning, albeit inversely, with linear neurons and a nonlinear plasticity rule (Hyvarinen and Oja, 2000). For variants of ICA based on information maximization (Bell and Sejnowski, 1997) or kurtosis (Hyvarinen and Oja, 2000) different nonlinearities arise (Fig. 2.2a), but Eq. 2.2 applies equally well. Hence, various instantiations of sparse coding and ICA models not only relate to each other in their normative assumptions (Olshausen and Field, 1997), but when implemented as iterative gradient update rules, they all employ nonlinear Hebbian learning.

2.2.3 Simple cell development for a large class of nonlinearities

Since the models described above can be implemented by similar plasticity rules, we hypothesized nonlinear Hebbian learning to be a general principle that explains the development of receptive field selectivity. Nonlinear Hebbian learning with an effective nonlinearity f is linked to an optimization principle with a function $F(z) = \int_0^z f(x)dx$ (Oja et al., 1991; Fyfe and Baddeley, 1995). For an input ensemble $\mathbf{x}$, optimality is achieved by weights $\tilde{\mathbf{w}}$ that maximize $\langle F(\tilde{\mathbf{w}}^T\mathbf{x})\rangle$, where angular brackets denote the average over the input statistics. Nonlinear Hebbian learning is a stochastic gradient ascent implementation of this optimization process, called projection pursuit (Friedman, 1987; Oja et al., 1991; Fyfe and Baddeley, 1995):

$$\tilde{\mathbf{w}} = max_{\mathbf{w}}\langle F(\mathbf{w}^T\mathbf{x})\rangle \implies \Delta\mathbf{w} \propto \mathbf{x}\, f(\mathbf{w}^T\mathbf{x}) \tag{2.3}$$

Motivated by results from ICA theory (Hyvarinen and Oja, 1998) and statistical properties of whitened natural images (Field, 1994), we selected diverse Hebbian nonlinearities f (Fig. 2.2a) and calculated the corresponding optimization value $\langle F(\mathbf{w}^T\mathbf{x})\rangle$ for different features of interest that we consider as candidate RF shapes, with a whitened ensemble of patches extracted from natural images as input (see Methods). These include a random connectivity pattern, a non-local oriented edge (as in principal components of natural images) and localized oriented edges (as in cat and monkey simple cells in the visual cortex), shown in Fig. 2.2b. The relative value of $\langle F(\mathbf{w}^T\mathbf{x})\rangle$ between one feature and another was remarkably consistent across various choices of the nonlinearity f, with localized orientation-selective receptive fields as maxima (Fig. 2.2b). Furthermore, we also searched for the maxima through gradient ascent, so as to confirm that the maxima are orientation selective (Fig. 2.2c, left). Our results indicate that receptive field development of simple cells is mainly governed by the statistical properties of natural images, while robust to specific model assumptions.

The relevant property of natural image statistics is that the distribution of a feature derived from typical localized oriented patterns has high kurtosis (Field, 1994; Olshausen and Field, 1996; Ruderman and Bialek, 1994). Thus to establish a quantitative measure whether a nonlinearity is suitable for feature learning, we define a *selectivity index* (*SI*), which measures the relative value of $\langle F(.)\rangle$ between a variable l with a Laplacian distribution and a variable g with Gaussian distribution (Hyvarinen and Oja, 1998): $SI = (\langle F(l)\rangle - \langle F(g)\rangle)/\sigma_F$ (see Methods). The Laplacian variable has higher kurtosis than the Gaussian variable, serving as a prototype of a kurtotic distribution. Since values obtained by filtering natural images with localized oriented patterns have a distribution with longer tails than other patterns (Field, 1994), as does the Laplacian variable compared to the Gaussian, positive values $SI > 0$ indicate good candidate functions for learning simple cell-like receptive fields from natural images. We find that each model has an appropriate parameter range where $SI > 0$ (Fig. 2.3). For example the quadratic rectifier nonlinearity needs an LTP threshold θ_2 below some critical level, so as to be useful for feature learning (Fig. 2.3a).

A sigmoidal function with threshold at zero has *negative SI*, but a *negative* sigmoid, as used in ICA studies (Bell and Sejnowski, 1997), has $SI > 0$. More generally, whenever an effective nonlinearity f is not suited for feature learning, its opposite $-f$ should be, since its SI will have the opposite sign (Fig. 2.2c). This implies that, in general, half of the function space could be suitable for feature learning (Hyvarinen and Oja, 1998), i.e. it finds weights w such that the distribution of the feature $\mathbf{w}^T\mathbf{x}$ has a long tail, indicating high kurtosis ("kurtotic feature"). The other half of the function space learns the least kurtotic features (e.g. random connectivity patterns for natural images, Fig. 2.2b,c).

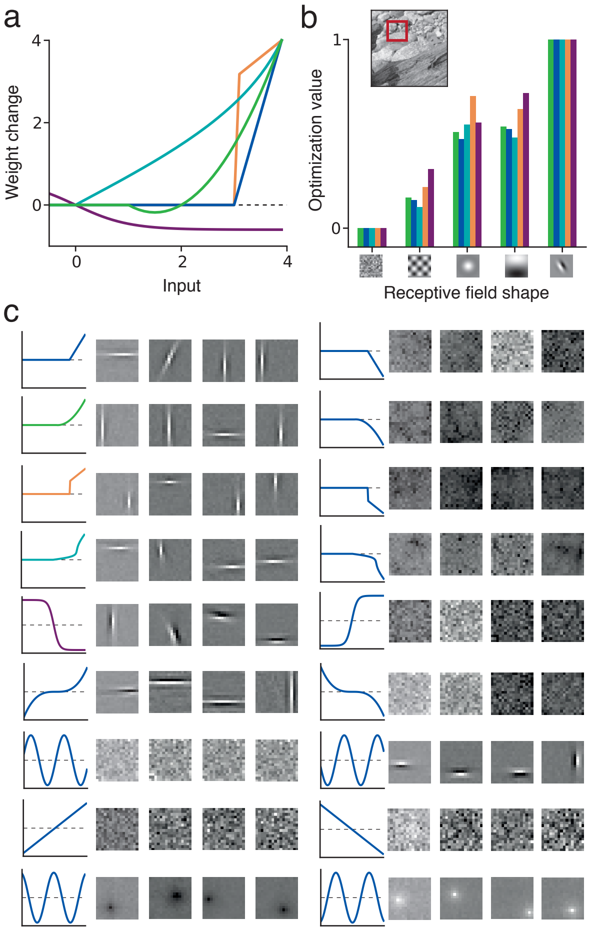

Figure 2.2 – **Simple cell development from natural images regardless of specific effective Hebbian nonlinearity.** (**a**) Effective nonlinearity of five common models (arbitrary units): quadratic rectifier (green, as in cortical and BCM models, $\theta_1 = 1.$, $\theta_2 = 2.$), linear rectifier (dark blue, as in L_1 sparse coding or networks with linear STDP, $\theta = 3.$), Cauchy sparse coding nonlinearity (light blue, $\lambda = 3.$), L_0 sparse coding nonlinearity (orange, $\lambda = 3.$), and negative sigmoid (purple, as in ICA models). (**b**) Relative optimization value $\langle F(\mathbf{w}^T\mathbf{x})\rangle$ for each of the five models in **a**, for different preselected features **w**, averaged over natural image patches **x**. Candidate features are represented as two-dimensional receptive fields. For all models, the optimum is achieved at the localized oriented receptive field. Inset: Example of natural image and image patch (red square) used as sensory input. (**c**) Receptive fields learned in four trials for ten effective Hebbian functions f (from top: the five functions considered above, u^3, $-sin(u)$, u, $(|u|-2)_+$, $-cos(u)$) (**left column**), and their opposites $-f$ (**right column**). The first seven functions (above the dashed line) lead to localized oriented filters, while a sign-flip leads to random patterns. Linear or symmetric functions are exceptions and do not develop oriented filters (**bottom rows**).

This universality strongly constrains the possible shape of receptive fields that may arise during development for a given input dataset. For whitened natural images, a learnable receptive field is in general either a localized edge detector or a non-localized random connectivity pattern.

An important special case is an effective linear curve, $f(u) = u$, which arises when both f-I and plasticity curves are linear (Miller et al., 1989). Because the linear model maximizes variance $\langle(\mathbf{w}^T\mathbf{x})^2\rangle$, it can perform principal component analysis (Oja, 1982), but does not have any feature selectivity on whitened input datasets, where variance is constant (Fig. 2.2c).

Symmetric effective nonlinearities, $f(u) = f(-u)$, are also exceptions, since their corresponding optimization functions are asymmetric, $F(u) = -F(-u)$, so that for datasets with symmetric statistical distributions, $P(\mathbf{x}) = P(-\mathbf{x})$, the optimization value will be zero, $\langle F_{asym.}(\mathbf{w}^T\mathbf{x}_{sym.})\rangle = 0$. As natural images are not completely symmetric, localized receptive fields do develop, though without orientation selectivity, as illustrated by a cosine function and a symmetric piece-wise linear function as effective nonlinearities (Fig. 2.2c, bottom rows).

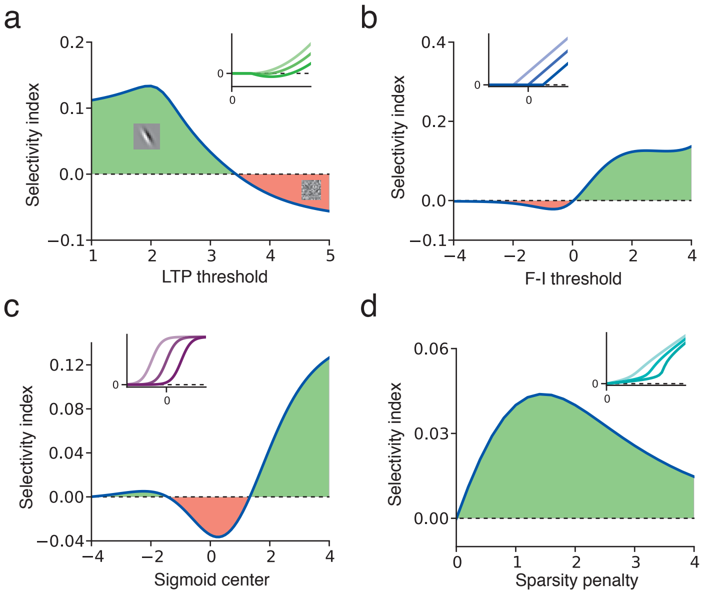

Figure 2.3 – **Selectivity index for different effective nonlinearities.** (**a**) Quadratic rectifier (small graphic, three examples with different LTP thresholds) with LTD threshold at $\theta_1 = 1$: LTP threshold must be below 3.5 to secure positive selectivity index (green region, main Fig) and learn localized oriented receptive fields (inset). A negative selectivity index (red region) leads to a random connectivity pattern (inset) (**b**) Linear rectifier: activation threshold must be above zero. (**c**) Sigmoid: center must be below $a = -1.2$ or, for a stronger effect, above $a = +1.2$. The opposite conditions apply to the negative sigmoid. (**d**) Cauchy sparse coding nonlinearity: positive but weak feature selectivity for any sparseness penalty $\lambda > 0$. Insets show the nonlinearities for different choices of parameters.

2.2.4 Receptive field diversity

Sensory neurons display a variety of receptive field shapes (Ringach, 2002), and recent modeling efforts (Rehn and Sommer, 2007; Zylberberg et al., 2011) have attempted to understand the properties that give rise to the specific receptive fields seen in experiments. We show here that the shape diversity of a model can be predicted by our projection pursuit analysis, and is primarily determined by the statistics of input representation, while relatively robust to the specific effective nonlinearity.

We studied a model with multiple neurons in the second layer, which compete with each other for the representation of specific features of the input. Each neuron had a piece-wise linear f-I curve and a quadratic rectifier plasticity function (see Methods) and projected inhibitory connections v onto all others. These inhibitory connections are learned by anti-Hebbian plasticity and enforce decorrelation of neurons, so that receptive fields represent different positions, orientations and shapes (Foldiak, 1990; Vogels et al., 2011; King et al., 2013). For 50 neurons, the resulting receptive fields became diversified (Fig. 2.4a-c, colored dots). In an overcomplete network of 1000 neurons, the diversity further increased (Fig. 2.4d-f, colored dots).

For the analysis of the simulation results, we refined our inspection of optimal oriented receptive fields for natural images by numerical evaluation of the optimality criterion $\langle F(\mathbf{w}^T\mathbf{x})\rangle$ for receptive fields $\mathbf{w} = \mathbf{w}_{Gabor}$, described as Gabor functions of variable length, width and spatial frequency. For all tested nonlinearities, the optimization function for single-neuron receptive fields varies smoothly with these parameters (Fig 2.4, grey-shaded background). The single-neuron optimality landscape was then used to analyze the multi-neuron simulation results. We found that receptive fields are located in the area where the single-neuron optimality criterion is near its maximum, but spread out so as to represent different features of the input (Fig. 2.4). Thus the map of optimization values, calculated from the theory of effective nonlinearity, enables us to qualitatively predict the shape diversity of receptive fields.

Although qualitatively similar, there are differences in the receptive fields developed for each model, such as smaller lengths for the L_0 sparse coding model (Fig. 2.4c). While potentially significant, these differences across models may be overwhelmed by differences due to other model properties, including different network sizes or input representations. This is illustrated by observing that receptive field diversity for a given model differ substantially across network sizes (Fig. 2.4), and the difference is even greater from simulations with an input that is not completely white (Fig. 2.5c). Thus our results suggests that efforts to model receptive field shapes observed experimentally (Ringach, 2002; Rehn and Sommer, 2007; Zylberberg et al., 2011) should focus on network size and input representation, which potentially have a stronger effect than the nonlinear properties of the specific model under consideration.

We also studied the variation of receptive field position and orientation. For all five nonlinearities considered, the optimization value is equal for different positions of the receptive field centers, confirming the translation invariance in the image statistics, as long as the receptive field is not too close to the border of the anatomically allowed fan-in of synaptic connections (Fig. 2.6b). Also, all nonlinearities reveal the same bias towards the horizontal and vertical orientations (Fig. 2.6c). These optimality predictions are confirmed in single neuron simu-

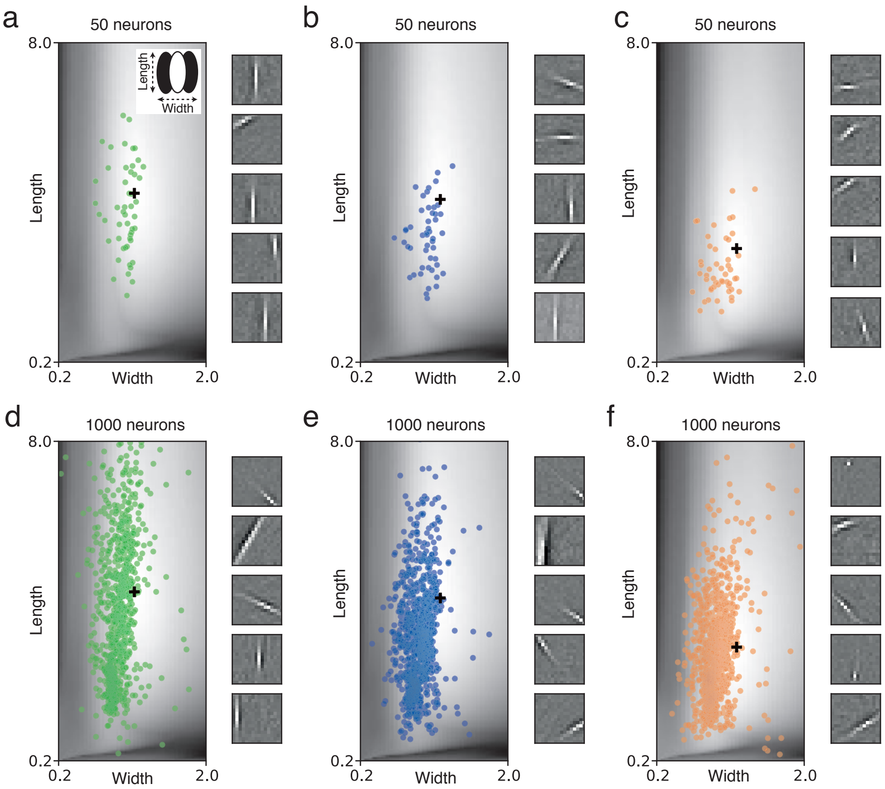

Figure 2.4 – **Optimal receptive field shapes in model networks induce diversity.** (**a-f**) Gray level indicates the optimization value for different lengths and widths (see inset in **a**) of oriented receptive fields for natural images, for the quadratic rectifier (left, see Fig. 2.2a), linear rectifier (middle) and L_0 sparse coding (right). Optima marked with a black cross. (**a-c**) Colored circles indicate the receptive fields of different shapes developed in a network of 50 neurons with lateral inhibitory connections. Insets on the right show example receptive fields developed during simulation. (**d-f**) Same for a network of 1000 neurons.

lations, which lead mostly to either horizontal or vertical orientations, at random positions (Fig. 2.6d). When the network is expanded to 50 neurons, recurrent inhibition forces receptive fields to cover different positions, though excluding border positions, and some neurons have non-cardinal orientations (Fig. 2.6e). With 1000 neurons, receptive fields diversify to many possible combinations of position, orientation and length (Fig. 2.6f).

2.2.5 Beyond V1 simple cells

Nonlinear Hebbian learning is not limited to explaining simple cells in V1. We investigated if the same learning principles apply to receptive field development in other visual or auditory areas or under different rearing conditions.

For auditory neurons (Smith and Lewicki, 2006), we used segments of speech as input (Fig. 2.7a) and observed the development of spectrotemporal receptive fields localized in both

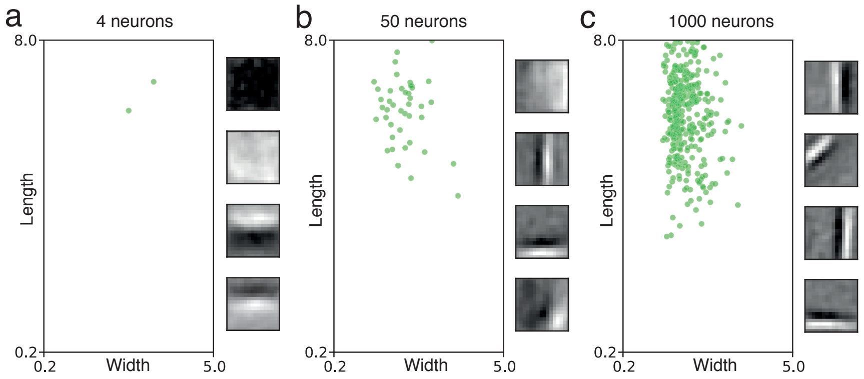

Figure 2.5 – **Receptive fields for non-whitened natural images.**Images were preprocessed as in the original sparse coding study (Olshausen and Field, 1997). We simulated linear rectifier neurons ($\theta = 0.5$) with a quadratic plasticity nonlinearity ($b = 0.5$). (**a**) Multiple-neuron simulations, with 4 neurons. The principal components dominate the optimization and receptive fields are not local, since they extend over most of the image patch. With 50 (**b**) and 1000 (**c**) neurons, lateral inhibition promotes diversity, and more localized receptive field are formed. (**insets**) Sample receptive fields developed for each simulation.

frequency and time (Miller et al., 2002) (Fig. 2.7d). The statistical distribution of input patterns aligned with the learned receptive fields had longer tails than for random or non-local receptive fields, indicating temporal sparsity of responses (Fig. 2.7d). Similar to our simple cell results, the learned receptive fields show higher optimization value for all five effective nonlinearities (Fig 2.7g).

For a study of receptive field development in the secondary visual cortex (V2) (Lee et al., 2007), we used natural images and the standard energy model (Hyvarinen et al., 2009) of V1 complex cells to generate input to V2 (Fig. 2.7b). The learned receptive field was selective to a single orientation over neighboring positions, indicating a higher level of translation invariance. When inputs were processed with this receptive field, we found longer tails in the feature distribution than with random features or receptive fields without orientation coherence (Fig 2.7e), and the learned receptive field had a higher optimization value for all choices of nonlinearity (Fig 2.7h).

Another important constraint for developmental models are characteristic deviations, such as strabismus, caused by abnormal sensory rearing. Under normal binocular rearing conditions, the fan-in of synaptic input from the left and right eyes overlap in visual space (Fig 2.7c). In this case, binocular receptive fields with similar features for left and right eyes develop. In the strabismic condition, the left and right eyes are not aligned, modeled as binocular rearing with non-overlapping input from each eye (Fig. 2.7c). In this scenario, a monocular simple cell-like receptive field developed (Fig. 2.7f), as observed in experiments and earlier models (Cooper et al., 2004). The statistical distributions confirm that for disparate inputs the monocular receptive field is more kurtotic than a binocular one, explaining its formation in diverse models (Hunt et al., 2013) (Fig 2.7f,i).

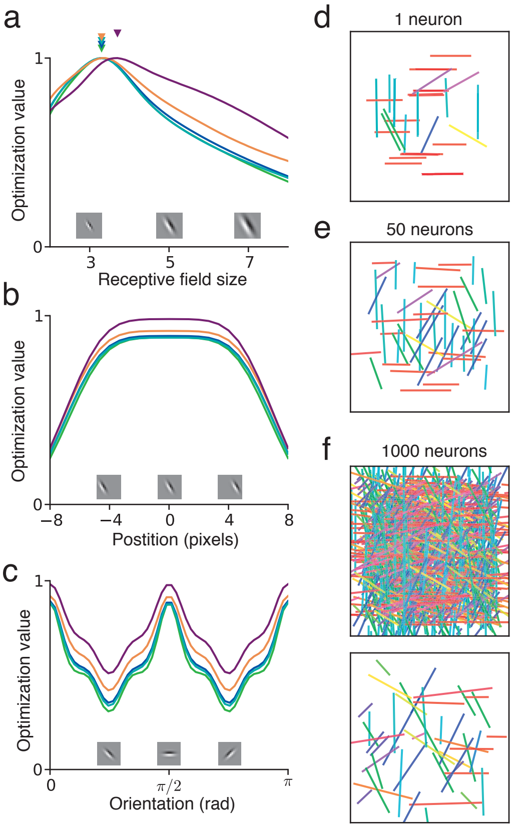

Figure 2.6 – **Diversity of receptive field size, position and orientation.** (**a**) The optimization value of localized oriented receptive fields, within a 16x16 pixel patch of sensors, as a function of size (see Methods), for five nonlinearities (colors as in Fig. 2.2a). Optimal size is a receptive field of width around 3 to 4 pixels (filled triangles). (**b**) The optimization value as a function of position of the receptive field center, for a receptive field width of 4 pixels, indicates invariance to position within the 16x16 patch, except near the borders. (**c**) The optimization value as a function of orientation shows preference toward horizontal and vertical directions, for all five nonlinearities. (**d**) Receptive field position, orientation and length (colored bars) learned for 50 single-neuron trials. The color code indicates different orientations. (**e**) Receptive field positions and orientations learned in a 50 neuron network reveal diversification of positions, except at the borders. (**f**) With 1000 neurons, positions and orientations cover the full range of combinations (top). Selecting 50 randomly chosen receptive fields highlights the diversification of position, orientation and size (bottom). Receptive fields were learned through the quadratic rectifier nonlinearity ($\theta_1 = 1.$, $\theta_2 = 2.$).

Our results demonstrate the generality of the theory across multiple cortical areas. Selecting a relevant feature space for an extensive analysis, as we have done with simple cells and natural images, may not be possible in general. Nonetheless, nonlinear Hebbian learning helps to explain why some features (and not others) are learnable in network models (Saxe et al., 2011).

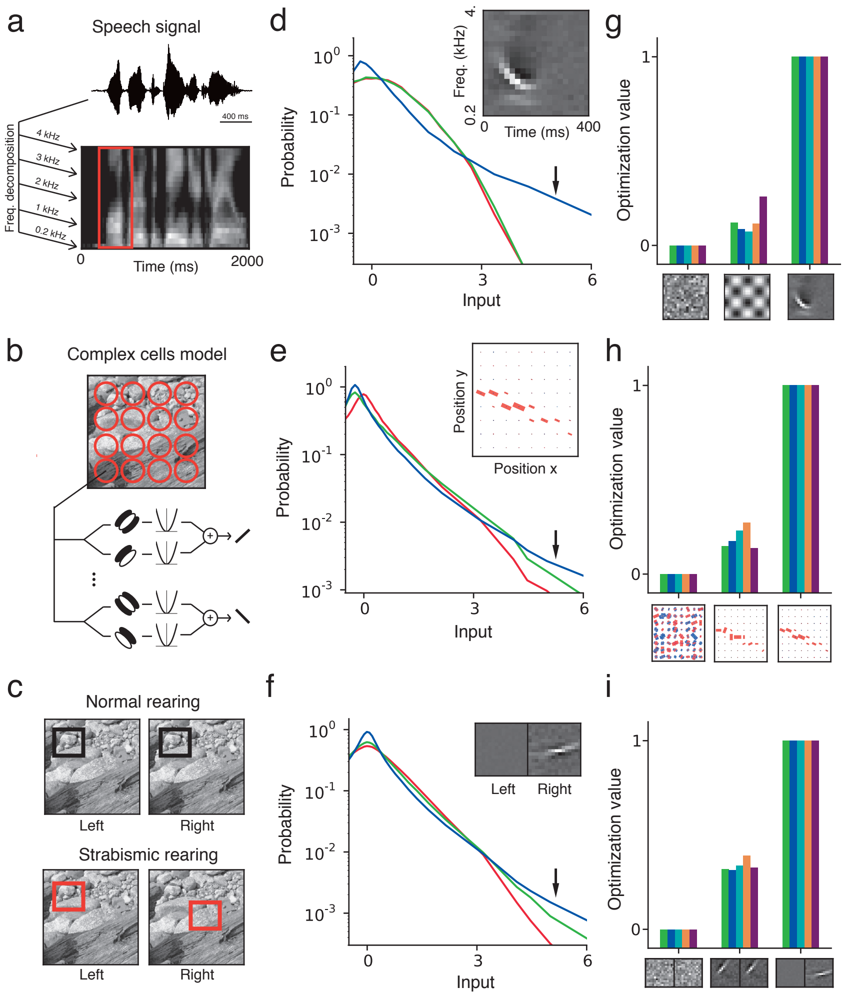

Figure 2.7 – **Nonlinear Hebbian learning across sensory modalities.** (**a**) The auditory input is modeled as segments over time and frequency (red) of the spectrotemporal representation of speech signals. (**b**) The V2 input is assembled from the output of modeled V1 complex cells at different positions and orientations. Receptive fields are represented by bars with size proportional to the connection strength to the complex cell with the respective position and orientation. (**c**) Strabismic rearing is modeled as binocular stimuli with non-overlapping left and right eye input patches (red). (**d-f**) Statistical distribution (log scale) of the input projected onto three different features for speech (**d**), V2 (**e**) and strabismus (**f**). In all three cases, the learned receptive field (blue, inset) is characterized by a longer tailed distribution (arrows) than the random (red) and comparative (green) features. (**g-i**) Relative optimization value for five nonlinearities (same as in Fig. 2.2), for the three selected patterns (**insets**). The receptive fields learned with the quadratic rectifier nonlinearity ($\theta_1 = 1.$, $\theta_2 = 2.$) are the maxima among the three patterns, for all five nonlinearities, for all three datasets.

2.3 Discussion

Historically, a variety of models have been proposed to explain the development and distribution of receptive fields. We have shown that nonlinear Hebbian learning is a parsimonious principle which is implicitly or explicitly present in many developmental models (Olshausen and Field, 1996; Bell and Sejnowski, 1997; Law and Cooper, 1994; Rehn and Sommer, 2007; Clopath et al., 2010; Savin et al., 2010; Zylberberg et al., 2011; Pfister and Gerstner, 2006; Hyvarinen and Oja, 1998; Foldiak, 1990; Hunt et al., 2013). The fact that receptive field development is robust to the specific nonlinearity highlights a functional relation between different models. It also unifies feature learning across sensory modalities: receptive fields form around features with a long-tailed distribution.

2.3.1 Relation to previous studies

Earlier studies have already placed developmental models side by side, comparing their normative assumptions, algorithmic implementation or receptive fields developed. Though consistent with their findings, our results lead to revised interpretations and predictions.

The similarities between sparse coding and ICA are clear from their normative correspondence (Olshausen and Field, 1997). Nevertheless, the additional constraint in ICA, of having at most as many features as inputs, makes it an easier problem to solve, allowing for a range of suitable algorithms (Hyvarinen and Oja, 2000). These differ from algorithms derived for sparse coding, in which the inference step is difficult due to overcompleteness. We have shown that regardless of the specific normative assumptions, it is the common implementation of nonlinear Hebbian learning that explains similarities in their learning properties.

In contrast to the idea that in sparse coding algorithms overcompleteness is required for development of localized oriented edges (Olshausen and Field, 1997), we have demonstrated that a sparse coding model with a single neuron is mathematically equivalent to nonlinear Hebbian learning and learns localized filters in a setting that is clearly "undercomplete". Thus differences observed in receptive field shapes between sparse coding and ICA models (Ringach, 2002) are likely due to differences in network size and input preprocessing. For instance, the original sparse coding model (Olshausen and Field, 1997) applied a preprocessing filter that did not completely whiten the input, leading to larger receptive fields (Fig. 2.5).

Studies that derive spiking models from normative theories often interpret the development of oriented receptive fields as a consequence of its normative assumptions (Savin et al., 2010; Zylberberg et al., 2011). In a recent example, a spiking network has been related to the sparse coding model (Zylberberg et al., 2011), using neural properties defined ad hoc. Our results suggest that many other choices of neural activations would have given qualitatively similar receptive fields, independent of the sparse coding assumption. While in sparse coding the effective nonlinearity derives from a linear plasticity rule combined with a nonlinear f-I curve, our results indicate that a nonlinear plasticity rule combined with a linear neuron model would give the same outcome.

In order to distinguish between different normative assumptions, or particular neural implementations, the observation of "oriented filters" is not sufficient and additional constraints are needed. Similarly receptive shape diversity, another important experimental constraint, should also be considered with care, since it cannot easily distinguish between models either. Studies that confront the receptive field diversity of a model to experimental data (Rehn and Sommer, 2007; Zylberberg et al., 2011; Ringach, 2002) should also take into account input preprocessing choices and how the shape changes with an increasing network size, since we have observed that these aspects may have a larger effect on receptive field shape than the particulars of the learning model.

Empirical studies of alternative datasets, including abnormal visual rearing (Hunt et al., 2013), tactile and auditory stimuli (Saxe et al., 2011), have also observed that different unsupervised learning algorithms lead to comparable receptive fields shapes. Our results offer a plausible theoretical explanation for these findings.

Past investigations on nonlinear Hebbian learning (Fyfe and Baddeley, 1995; Hyvarinen and Oja, 1998) demonstrated that many nonlinearities were capable of solving the cocktail party problem. Since it is a specific toy model, that asks for the unmixing of linearly mixed independent features, it is not clear a priori whether the same conclusions would hold in other settings. We have shown that the results of Fyfe and Baddeley (1995) and Hyvarinen and Oja (1998) generalize in two directions. First, the effective nonlinear Hebbian learning mechanism is also behind other models beyond ICA, such as sparse coding models and plastic spiking networks. Second, the robustness to the choice of nonlinearity is not limited to a toy example, but also holds in multiple real world data. Together, these insights explain and predict the outcome of many developmental models, in diverse applications.

2.3.2 Robustness to normative assumptions

Many theoretical studies start from normative assumptions (Bell and Sejnowski, 1997; Rehn and Sommer, 2007; Savin et al., 2010; Olshausen and Field, 1997), such as a statistical model of the sensory input or a functional objective, and derive neural and synaptic dynamics from them. Our claim of universality of feature learning indicates that details of normative assumptions may be of lower importance.

For instance, in sparse coding one assumes features with a specific statistical prior (Rehn and Sommer, 2007; Olshausen and Field, 1997). After learning, this prior is expected to match the posterior distribution of the neuron's firing activity (Rehn and Sommer, 2007; Olshausen and Field, 1997). Nevertheless, we have shown that receptive field learning is largely unaffected by the choice of prior. Thus, one cannot claim that the features were learned because they match the assumed prior distribution, and indeed in general they do not. For a coherent statistical interpretation, one could search for a prior that would match the feature statistics. However, since the outcome of learning is largely unaffected by the choice of prior, such a statistical approach would have limited predictive power. Generally, kurtotic prior assumptions enable feature learning, but the specific priors are not as decisive as one might expect. Because normative approaches have assumptions, such as independence of hidden features, that are not generally satisfied by the data they are applied to, the actual algorithm that is used for optimization becomes more critical than the formal statistical model.

The concept of sparseness of neural activity is used with two distinct meanings. The first one is a single-neuron concept and specifically refers to the long-tailed distribution statistics of neural activity, indicating a "kurtotic" distribution. The second notion of sparseness is an ensemble concept and refers to the very low firing rate of neurons, observed in cortical activity (Barth and Poulet, 2012), which may arise from lateral competition in overcomplete representations. Overcompleteness of ensembles makes sparse coding different from ICA (Olshausen and Field, 1997). We have shown here that competition between multiple neurons is fundamental for receptive field diversity, whereas it is not required for simple cell formation per se. Kurtotic features can be learned even by a single neuron with nonlinear Hebbian learning, and with no restrictions on the sparseness of its firing activity.

2.3.3 Interaction of selectivity with preprocessing and homeostasis

The concept of nonlinear Hebbian learning also clarifies the interaction of feature selectivity with preprocessing mechanisms. We have assumed whitened data throughout the study, except Fig. 2.5. Since after whitening second-order correlations are uninformative, neurons can develop sensitivity to higher order features. While whitened data is formally not required for our analysis, second-order correlations may dominate the optimization for non-white input, so that principal components will be learned (Fig. 2.5a). Only when multiple neurons are added and receptive fields diversify, are localized simple cells formed with an input that is not completely white (Olshausen and Field, 1997) (Fig. 2.5c).

In studies of spiking networks, the input is restricted to positive rates, possibly through an on/off representation, as observed in the LGN (Miller, 1994). While the center-surround properties of LGN contributes to a partial decorrelation of neuronal activity (Dan et al., 1996), in such alternative representations, trivial receptive fields may develop, such as a single non-zero synapse, and additional mechanisms, such as hard bounds on each synaptic strength, $a \leq w_j \leq b$, may be necessary to restrict the optimization space to desirable features (Clopath et al., 2010).

Instead of constraining the synaptic weights, one may implement a synaptic decay as in Oja's plasticity rule (Oja, 1982), $\Delta w \propto x \cdot y - w \cdot y^2$ (see also (Chen et al., 2013)). Because of its multiplicative effect, the decay term does not alter the receptive field, but only scales its strength. Thus, it is equivalent to rescaling the input in the f-I curve, so as to shift it to the appropriate range (Fig. 2.3). Similar scaling effects arise from f-I changes due to intrinsic plasticity (Savin et al., 2010; Turrigiano, 2011; Elliott, 2014). The precise relation between nonlinear Hebbian learning, spiking representations and homeostasis in the cortex is an important topic for further studies.

2.3.4 Universality supports biological instantiation

The principle of nonlinear Hebbian learning has a direct correspondence to biological neurons and is compatible with a large variety of plasticity mechanisms. It is not uncommon for biological systems to have diverse implementations with comparable functional properties (Prinz et al., 2004). Different species, or brain areas, could have different neural and plasticity characteristics, and still have similar feature learning properties (Sharma et al., 2000; Kaschube et al., 2010). The generality of the results discussed in this paper reveals learning simple cell-like receptive fields from natural images to be much easier than previously thought. It implies that a biological interpretation of models is possible even if some aspects of a model appear simplified or even wrong in some biological aspects. Universality also implies that the study of receptive field development is not sufficient to distinguish between different models.

The relation of nonlinear Hebbian learning to projection pursuit endorses the interpretation of cortical plasticity as an optimization process. Under the rate coding assumptions considered here, the crucial property is an effective synaptic change linear in the pre-synaptic rate, and nonlinear in the post-synaptic input. Pairing experiments with random firing and independently varying pre- and post-synaptic rates would be valuable to investigate these properties (Sjostrom et al., 2001, 2008; Graupner and Brunel, 2012). Altogether, the robustness to details in both input modality and neural implementation suggests nonlinear Hebbian learning as a fundamental principle underlying the development of sensory representations.

2.4 Methods

Spiking model. A generalized leaky integrate-and-fire neuron (Pozzorini et al., 2013) was used as spiking model, which includes power-law spike-triggered adaptation and stochastic firing, with parameters (Pozzorini et al., 2013) fitted to pyramidal neurons. The f-I curve $g(I)$ was estimated by injecting step currents and calculating the trial average of the spike count over the first 500 ms. The minimal triplet-STDP model(Pfister and Gerstner, 2006) was implemented, in which synaptic changes follow

$$\frac{d}{dt}w(t) = A^{+}y(t)\bar{y}^{+}(t)\bar{x}^{+}(t) - A^{-}x(t)\bar{y}^{-}(t) \tag{2.4}$$

where $y(t)$ and $x(t)$ are the post- and pre-synaptic spike trains, respectively: $y(t) = \sum_f \delta(t - t^f)$, where t^f are the firing times and δ denotes the Dirac δ-function; $x(t)$ is a vector with components $x_i(t) = \sum_f \delta(t - t_i^f)$, where t_i^f are the firing times of pre-synaptic neuron i; w is a vector comprising the synaptic weights w_i connecting a pre-synaptic neuron i to a post-synaptic cell. $A^+ = 6.5 \cdot 10^{-3}$ and $A^- = 5.3 \cdot 10^{-3}$ are constants, and $\bar{y}^+$, $\bar{x}^+$ and $\bar{y}^-$ are moving averages, implemented by integration (e.g. $\tau\frac{\partial \bar{y}}{\partial t} = -\bar{y} + y$), with time scales 114.0 ms, 16.8 ms and 33.7 ms, respectively (Pfister and Gerstner, 2006). For estimating the nonlinearity $h(y)$ of the plasticity, pre- and post-synaptic spike trains were generated as Poisson processes, with the pre-synaptic rate set to 20 Hz.

A linear rectifier $g(x) = a(x-b)_+$ was fitted to the f-I curve of the spiking neuron model by squared error optimization. Similarly, a quadratic function $h(x) = a(x^2 - bx)$ was fitted to the nonlinearity of the triplet STDP model. The combination of these two fitted functions was plotted as fit for the effective nonlinearity $f(x) = h(g(x))$.

Sparse coding analysis. A sparse coding model, with K neurons $y_1, \ldots, y_K$, has a nonlinear Hebbian learning formulation. The sparse coding model minimizes a least square reconstruction error between the vector of inputs $\mathbf{x}$ and the reconstruction vector $\mathbf{Wy}$, where $\mathbf{W} = [\mathbf{w}_1 \ldots \mathbf{w}_K]$, and $\mathbf{y} = (y_1, \ldots, y_K)$ is the vector of neuronal activities, with $y_j \geq 0$ for $1 \leq j \leq K$. The total error E combines a sparsity constraint S with weight λ and the reconstruction error, $E = \frac{1}{2}||\mathbf{x} - \mathbf{Wy}||^2 + \lambda \sum S(y_k)$. E has to be minimal, averaged across all input samples, under the constraint $y_j \geq 0$ for all j.

The minimization problem is solved by a two-step procedure. In the first step, for each input sample, one minimizes E with respect to all hidden units y_j

$$\begin{aligned}
\frac{d}{dy_j} E = 0 &\iff \mathbf{w}_j^T(\mathbf{x} - \mathbf{Wy}) - \lambda S'(y_j) = 0 \\
&\iff \mathbf{w}_j^T \mathbf{x} - \sum_{k \neq j} (\mathbf{w}_j^T \mathbf{w}_k) y_k - |\mathbf{w}_j|^2 y_j - \lambda S'(y_j) = 0 \\
&\iff y_j + \lambda S'(y_j) = \mathbf{w}_j^T \mathbf{x} - \sum_{k \neq j} (\mathbf{w}_j^T \mathbf{w}_k) y_k \\
&\iff y_j = g(\mathbf{w}_j^T \mathbf{x} - \sum_{k \neq j} v_{jk} y_k)
\end{aligned} \tag{2.5}$$

where we constrained the vector $\mathbf{w}_j$ of synapses projecting onto unit y_j by $|\mathbf{w}_j|^2 = 1$, defined the activation function $g(.) = T^{-1}(.)$, the inverse of $T(y) = (y + \lambda S'(y))$, and defined recurrent synaptic weights $v_{jk} = \mathbf{w}_j^T \mathbf{w}_k$. For each input sample $\mathbf{x}$, this equation shall be iterated until convergence. The equation can be interpreted as a recurrent neural network, where each neuron has an activation function g, and the input is given by the sum of the feedforward drive $\mathbf{w}_j^T \mathbf{x}$ and a recurrent inhibition term $-\sum_{k \neq j} v_{jk} y_k$. To avoid instability, we implement a smooth membrane potential u_j, which has the same convergence point (Rozell et al., 2008)

$$\begin{aligned}
\tau_u \frac{d}{dt} u_j(t) &= -u_j(t) + (\mathbf{w}_j^T \mathbf{x} - \sum_{k \neq j} v_{jk} y_k(t)) \\
y_j(t) &= g(u_j(t))
\end{aligned} \tag{2.6}$$

initialized with $u_j(t) = 0$.

The second step is a standard gradient descent implementation of the least square regression optimization, leading to a learning rule

$$\Delta \mathbf{w}_j \propto \frac{d}{dw_j} E = (\mathbf{x} - \mathbf{W}^T \mathbf{y})\ y_j = \mathbf{x}\ y_j - \mathbf{w}_j\ y_j^2 - \sum_{k \neq j} \mathbf{w}_k y_k y_j$$

The decay term $\mathbf{w}_j\ y_j^2$ has no effect, since the norm is constrained to $|\mathbf{w}_j| = 1$ at each step. For a single unit y, the model simplifies to a nonlinear Hebbian formulation, $\Delta \mathbf{w} \propto \mathbf{x}\ g(\mathbf{w}_j^T \mathbf{x})$. For multiple units, it can be interpreted as projection pursuit on an effective input, not yet represented by other neurons, $\tilde{\mathbf{x}_j} = \mathbf{x} - \sum_{k \neq j} \mathbf{w}_k y_k$, which simplifies to $\Delta \mathbf{w}_j \propto \tilde{\mathbf{x}}_j \cdot g(\mathbf{w}_j^T \tilde{\mathbf{x}_j})$.

There are two non-local terms that need to be implemented by local mechanisms so as to be biologically plausible. First, the recurrent weights depend on the overlap between receptive fields, $\mathbf{w}_j^T \mathbf{w}_k$, which is non-local. The sparse coding model assumes independent hidden neurons, which implies that after learning neurons should be pair-wise uncorrelated, $cov(y_j, y_k) = 0$. As an aside we note that the choice $v_{jk} = \mathbf{w}_j^T \mathbf{w}_k$ does not automatically guarantee decorrelation. Decorrelation may be enforced through plastic lateral connections, following an anti-Hebbian rule (Foldiak, 1990; Zylberberg et al., 2011), $\Delta v_{jk} \propto (y_j - \langle y_j \rangle) \cdot y_k$, where $\langle y_j \rangle$ is a moving average (we use $\tau = 1000$ input samples). Thus by substituting fixed recurrent connections by anti-Hebbian plasticity, convergence $\Delta v_{jk} = 0$ implies $cov(y_j, y_k) = 0$. While this implementation does not guarantee $v_{jk} = \mathbf{w}_j^T \mathbf{w}_k$ after convergence, neither does $v_{jk} = \mathbf{w}_j^T \mathbf{w}_k$ guarantee decorrelation $cov(y_j, y_k) = 0$, it does lead to optimal decorrelation, which is the basis of the normative assumption. Additionally we constrain $v_{jk} \geq 0$ to satisfy Dale's law. Although some weights would converge to negative values otherwise, most neuron pairs have correlated receptive fields, and thus positive recurrent weights.

Second, we ignore the non-local term $\sum_{k \neq j} \mathbf{w}_k y_k y_j$ in the update rule. Although this approximation is not theoretically justified, we observed in simulations that receptive fields do not qualitatively differ when this term is removed.

The resulting Hebbian formulation can be summarized as

$$\begin{aligned} y_j &= g(\mathbf{w}_j^T \mathbf{x} - \sum_{k \neq j} v_{jk} y_k) \\ \Delta \mathbf{w}_j &\propto \mathbf{x}\ y_j \\ \Delta v_{jk} &\propto (y_j - \langle y_j \rangle) \cdot y_k \end{aligned} \tag{2.7}$$

This derivation unifies previous results on the biological implementation of sparse coding: the relation of the sparseness constraint to a specific activation function (Rozell et al., 2008), the derivation of a Hebbian learning rule from quadratic error minimization (Oja, 1982), and the possibility of approximating lateral interaction terms by learned lateral inhibition (Foldiak, 1990; Zylberberg et al., 2011).

Nonlinearities and optimization value. The optimization value for a given effective nonlinearity f, synaptic weights w, and input samples x, is given by $R = \langle F(\mathbf{w}^T\mathbf{x})\rangle$, where $F(z) = \int_0^z f(x)dx$ and angular brackets indicate the ensemble average over x. Relative optimization values in Figs. 2.2b and 2.6 were normalized to $[0,1]$, relative to the minimum and maximum values among the considered choice of features w, $R^* = (R - R_{min})/(R_{max} - R_{min})$. The selectivity index of a nonlinearity f is defined as $SI = (\langle F(l)\rangle - \langle F(g)\rangle)/\sigma_F$, where l and g are Laplacian and Gaussian variables respectively, normalized to unit variance. $\sigma_F = \sqrt{\sigma_{F(l)}\sigma_{F(g)}}$ is a normalization factor, with $\sigma_{F(.)} = \sqrt{\langle F(.)^2\rangle}$. The selectivity of an effective nonlinearity f is not altered by multiplicative scaling, $\tilde{f}(u) = \alpha f(u)$, neither by additive constants when the input distribution is symmetric, $\tilde{f}(u) = \alpha f(u) + \beta$. The effective nonlinearities in Fig. 2.2 included the linear rectifier $f(u) = \begin{cases} 0, & if\ u < \theta \\ u - \theta, & if\ u \geq \theta \end{cases}$, the quadratic rectifier $f(u) = \begin{cases} 0, & if\ u < \theta \\ (u-\theta)(u-\theta-b), & if\ u \geq \theta \end{cases}$, the L_0 sparse coding nonlinearity $f(u) = \begin{cases} 0, & if\ u < \lambda \\ u, & if\ u \geq \lambda \end{cases}$, the Cauchy sparse coding nonlinearity $f = T^{-1}$, where $T(y) = \begin{cases} 0, & if\ y < 0 \\ y + 2\lambda y/(1+y^2), & if\ y \geq 0 \end{cases}$, the negative sigmoid $f(u) = 1 - 2/(1 + e^{-2u})$, a polynomial function $f(u) = u^3$, trigonometric functions $sin(u)$ and $cos(u)$, a symmetric piece-wise linear function $f(u) = \begin{cases} 0, & if\ |u| < \theta \\ |u| - \theta, & if\ |u| \geq \theta \end{cases}$, as well as, for comparison, a linear function $f(u) = u$.

Receptive field learning. Natural image patches (16 by 16 pixel windows) were sampled from a standard dataset (Olshausen and Field, 1996) (10^6 patches). Patches were randomly rotated by ±90° degrees to avoid biases in orientation. The dataset was whitened by mean subtraction and a standard linear transformation $\mathbf{x}^* = \mathbf{M}\mathbf{x}$, where

$$\mathbf{M} = \mathbf{R}\mathbf{D}^{-1/2}\mathbf{R}^T \tag{2.8}$$

and $\langle \mathbf{x}\mathbf{x}^T\rangle = \mathbf{R}\mathbf{D}\mathbf{R}^T$ is the eigenvalue decomposition of the input correlation matrix. In Fig. 2.5, we used images preprocessed as in Olshausen and Field (1996), filtered in the spatial frequency domain by $M(f) = f\ e^{-(f/f_0)^4}$. The exponential factor is a low-pass filter that attenuates high-frequency spatial noise, with $f_0 = 200$ cycles per image. The linear factor f was designed to whiten the images by canceling the approximately $1/f$ power law spatial correlation observed in natural images (Ruderman and Bialek, 1994). But since the exponent of the power law for this particular dataset has an exponent closer to 1.2, the preprocessed images exhibit higher variance at lower spatial frequencies.

Synaptic weights were initialized randomly (normal distribution with zero mean) and, for an effective nonlinearity f, evolved through $\mathbf{w}_{k+1} = \mathbf{w}_k + \eta\,\mathbf{x}\,f(\mathbf{w}_k^T\mathbf{x}_k)$, for each input sample x_k, with a small learning rate η. We enforced normalized weights at each time step, $|\mathbf{w}|_2 = 1$, through multiplicative normalization, implicitly assuming rapid homeostatic mechanisms (Turrigiano, 2011; Zenke et al., 2013). For multiple neurons, the neural version of the sparse coding model described in Eq 2.7 was implemented. In Fig 2.4 and 2.5, the learned receptive fields were fitted to Gabor filters by least square optimization. Receptive fields with less than 0.6 variance explained were rejected (less than 5% of all receptive fields).

Receptive field selection. In Fig. 2.2b, the five selected candidate patterns are: random connectivity filter (weights sampled independently from the normal distribution with zero mean), high-frequency Fourier filter (with equal horizontal and vertical spatial periods, $T_x = T_y = 8$ pixels), difference of Gaussians filter ($\sigma_1 = 3.$, $\sigma_2 = 4.$), low-frequency Fourier filter ($T_x = 16$, $T_y = 32$), and centered localized Gabor filter ($\sigma_x = 1.5$, $\sigma_y = 2.0$, $f = 0.2$, $\theta = \pi/3$, $\phi = \pi/2$). Fourier filters were modeled as $w_{ab} = sin(2\pi a/T_x) * cos(2\pi b/T_y)$; difference of Gaussians filters as the difference between two centered 2D Gaussians with same amplitude and standard deviations σ_1 and σ_2; and we considered standard Gabor filters, with center (x_c, y_c), spatial frequency f, width σ_x, length σ_y, phase ϕ and angle θ. In Fig 2.4 and 2.5 we define the Gabor width and length in pixels as 2.5 times the standard deviation of the respective Gaussian envelopes, σ_x and σ_y. In Fig. 2.6a, a Gabor filter of size s had parameters $\sigma_x = 0.3 \cdot s$, $\sigma_y = 0.6 \cdot s$, $f = 1/s$ and $\theta = \pi/3$. In Fig. 2.6b-c, the Gabor filter parameters were $\sigma_x = 1.2$, $\sigma_y = 2.4$, $f = 0.25$. All receptive fields were normalized to $|\mathbf{w}|_2 = 1$. In Fig. 2.4 and 2.5, the background optimization value was calculated for Gabor filters of different widths, lengths, frequencies, phases $\phi = 0$ and $\phi = \pi/2$. For each width and length, the maximum value among frequencies and phases was plotted.

Additional datasets. For the strabismus model, two independent natural image patches were concatenated, representing non-overlapping left and right eye inputs, forming a dataset with 16 by 32 patches (Cooper et al., 2004). For the binocular receptive field in the strabismus statistical analysis (Fig. 2.7a), a receptive field was learned with a binocular input with same input from left and right eyes. As V2 input, V1 complex cell responses were obtained from natural images as in standard energy models (Hyvarinen et al., 2009), modeled as the sum of the squared responses of simple cells with alternated phases. These simple cells were modeled as linear neurons with Gabor receptive fields ($\sigma_x = 1.2$, $\sigma_y = 2.4$, $f = 0.3$), with centers placed on a 8 by 8 grid (3.1 pixels spacing), with 8 different orientations at each position (total of 512 input dimensions). For the non-orientation selective receptive field in the V2 statistical analysis (Fig. 2.7d), the orientations of the input complex cells for the learned receptive field were randomized. As auditory input, spectrotemporal segments were sampled from utterances spoken by a US English male speaker (CMU US BDL ARCTIC database, Kominek and Black (2004)). For the frequency decomposition (Smith and Lewicki, 2006), each audio segment was filtered by gammatone kernels, absolute and log value taken and downsampled to 50 Hz. Each sample was 20 time points long (400 ms segment) and 20 frequency points wide (equally spaced between 0.2 kHz and 4.0 kHz). For the non-local receptive field in the auditory statistical analysis (Fig. 2.7g), a Fourier filter was used ($T_t = T_f = 10$). For all datasets, the input ensemble was whitened after the preprocessing steps, by the same linear transformation described above for natural images, and all receptive fields were normalized to $|\mathbf{w}|_2 = 1$.

3 Cortical synaptic plasticity as second-order invariant feature learning

Synaptic plasticity is believed to underlie cortical receptive field formation from natural input statistics. While nonlinear Hebbian potentiation can explain this development, it assumes artificially decorrelated inputs and stability constraints, typically attributed to depression and homeostatic mechanisms. Here we demonstrate how synaptic depression resolves both the limitation of decorrelation and stability. The linear anti-Hebbian character of synaptic depression is shown to make cortical plasticity invariant to second-order input statistics and explains receptive field development without the requirement of whitened inputs. It also provides robustness to heterogeneities in pre-synaptic firing rates and dendritic attenuation. These findings give a precise functional interpretation for synaptic potentiation, depression and homeostasis in cortical plasticity, which appears optimally designed for robust feature learning.

3.1 Introduction

Synaptic plasticity is believed to underlie cortical receptive field formation from natural input statistics. While we have shown in the previous chapter that nonlinear Hebbian potentiation (LTP) is sufficient to explain this development, we assumed artificially whitened inputs and artificial stability constraints on the norm of the synaptic weights.

The need to stabilize the weight dynamics in Hebbian models lead to the development of classic Hebbian models with an additional term for synaptic depression (LTD) or a homeostatic mechanism. In Oja's rule, synaptic depression has a supra-linear dependency on the neuron activity, avoiding run-away firing rates (Oja, 1989). BCM models include a supra-linear meta-plasticity factor that modulates synaptic depression, also avoiding firing rates to diverge (Bienenstock et al., 1982b). Any of these approaches has the desired stability effects, and it has not been clear how to differentiate between them functionally (Cooper et al., 2004).

Here, we demonstrate that the requirement to learn patterns from non-whitened inputs suggests a unifying theory for understanding synaptic depression and homeostasis in terms of feature learning. We show that the linear anti-Hebbian character of synaptic depression, observed in cortical plasticity experiments and BCM-like models, makes synaptic plasticity invariant to second-order input statistics. This invariance enables receptive field development without the requirement of whitened inputs. We also show that this invariance provides robustness to any network heterogeneities that correspond to linear transformations, such as diversity in pre-synaptic firing rates and dendritic attenuation of EPSP's.

Our findings give a precise functional interpretation for three central components of synaptic plasticity. Synaptic potentiation implements nonlinear Hebbian learning, while synaptic depression enables second-order invariance, and meta-plasticity maintains the precise balance between LTP and LTD required for second-order invariant learning. This provides support for the interpretation that cortical plasticity implements robust higher-order feature learning.

3.2 Results

3.2.1 Second-order invariant learning rules

As in the previous chapter, we define feature learning in terms of projection pursuit. For simplicity, we consider linear rectifier neurons, $y = (\mathbf{w}^T\mathbf{x})_+$, and the third order objective function, as it gives rises to the quadratic Hebbian nonlinearities observed in experiments (Pfister and Gerstner, 2006). The simplest case is the objective function

$$\mathbf{w}^* = \text{argmax}_{\mathbf{w},|\mathbf{w}|=1} \langle y^3 \rangle \tag{3.1}$$

As discussed in the previous chapter, if the input is not white, second-order statistics may dominate higher-order moments, and the neuron will develop receptive fields in the direction of highest variance, effectively implementing principal component analysis, and will not be sensitive to higher-order structure in the input.

We thus consider instead the normalized third-order moment as an objective function,

$$\mathbf{w}^* = \text{argmax}_{\mathbf{w}} \left\langle \left(\frac{y}{\sigma_y} \right)^3 \right\rangle \tag{3.2}$$

where $\sigma_y = \sqrt{\langle y^2 \rangle}$. The normalization factor makes the objective function scale invariant, and weights will not be driven to diverging norms as in the unnormalized case, allowing the weight norm $|\mathbf{w}|$ to be unconstrained. However, if unconstrained, the norm may slowly diffuse out of bounds, a problem we will address later on.

Such an objective function has the property of being invariant to linear transformations on the input, $\tilde{\mathbf{x}} = \mathbf{M}\,\mathbf{x}$ (see Methods). This linear invariance means that any linear transformation will be compensated by the optimal synaptic weights. As whitening is a linear transformation, the optimization problem for non-whitened inputs $\mathbf{x}$ will therefore give the same outcome as for its whitened version, and the results from the previous chapter will still hold.

By deriving the correspondent stochastic gradient ascent learning rule (see Methods), we arrive at

$$\Delta\mathbf{w} \propto \mathbf{x}\, y^2 - h^*(y)\, \mathbf{x}\, y \tag{3.3}$$

where $h^*(y) = \frac{\langle y^3\rangle}{\langle y^2\rangle}$ is referred to as the balancing homeostatic factor.

This learning rule is similar to the BCM model, though with a different homeostatic factor. Since this learning rule is scale invariant, the norm of the weights is unconstrained, so that given optimal weights $\mathbf{w}^*$, rescaled versions will also be optima, $\tilde{\mathbf{w}}^* = c\,\mathbf{w}^*$, for any positive constant c.

We confirm the expected properties of this learning rule in a two-dimensional toy model, in which the input, $\mathbf{x} = M\,\mathbf{s}$, is a linear combination of a Laplacian and a normal variable, $\mathbf{s} = [l_1, n_1]^T$, where the Laplacian variable is a proxy for a higher-order hidden feature. In Figure 3.1a, we show that the weights converge to the hidden feature, even though its variance is smaller than the variance in the other direction where the distribution is Gaussian. However, there is no constraint on the norm of $\mathbf{w}$, and all vectors in the direction of the hidden feature are optima.

We may verify that there is indeed no selectivity for the norm of the weights by calculating the temporal derivative of the norm $\left\langle \frac{\partial}{\partial t}|\mathbf{w}|^2 \right\rangle = \left\langle \mathbf{w}^T\, \frac{\Delta\mathbf{w}}{\Delta t} \right\rangle$, which is therefore equivalent to projecting the gradient in the direction of $\mathbf{w}$,

$$\langle \mathbf{w}^T \Delta\mathbf{w}\rangle \propto \langle \mathbf{w}^T\mathbf{x}y^2 - \frac{\langle y^3\rangle}{\langle y^2\rangle}\mathbf{w}^T\mathbf{x}y\rangle = \langle y^3\rangle - \frac{\langle y^3\rangle}{\langle y^2\rangle}\langle y^2\rangle = 0. \tag{3.4}$$

Thus, with stochastic gradient rule (Eq. 3.3), the norm stays on average stable. At most it may show a slow diffuse (drift-free) trajectory.

3.2.2 Stable second-order invariance

Instead of implementing the balancing homeostatic factor, $h^*(y)$, we are in fact free to choose any other homeostatic factor $h(y)$ in Eq. 3.3, as long as it stabilizes the norm of the weights. This perhaps anti-intuitive property follows from the fact that after convergence we have $\Delta\mathbf{w}^T\mathbf{e} = 0$ for any direction $\mathbf{e}$. Therefore, the potentiation and depression (c.f. Eq. 3.3) must cancel each other in the direction of weight vector, which implies that after convergence $h(y) = \frac{\langle y^3\rangle}{\langle y^2\rangle} = h^*(y)$ (see Methods). In general, any supra-linear homeostatic factor will be stable. An example is the homeostatic factor in the BCM model, $h(y) = \langle y^2\rangle$. Thus, whatever the stable homeostatic function $h(y)$, once the norm of the weights stabilizes, the learning process will be the same as for the model with $h^*(y)$.

In Figure 3.1b, we show that the stable model with the BCM homeostatic factor has the same

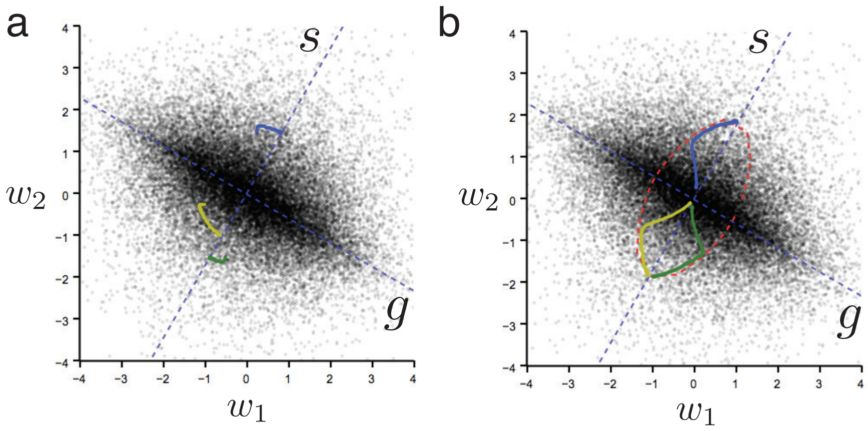

Figure 3.1 – **Second-order invariant learning in a two-dimensional toy model.** (**a**) For the original second-order invariant learning rule, the initial weights converge (colored paths) to the direction of the hidden feature (direction s), even though it has much smaller variance than other directions. Since there is no constraint on the norm of the weights, all vectors in the direction of the hidden feature are optima. Direction g is the direction of the normal variable. Grey dots on the background represent data points $\mathbf{x}_i$. (**b**) For the stable second-order invariant learning rule (BCM model), the weights converge to the direction of the hidden feature, with the norm stabilized by the homeostatic factor. The red dashed line represents the theoretically predicted value for the stable norm at each direction.

second-order invariant selectivity properties of the balanced version, but its norm is restricted to the values in which $h(y) = \frac{\langle y^3 \rangle}{\langle y^2 \rangle}$, which can be calculated analytically (see Methods).

3.2.3 Receptive field development with non-whitened inputs

Since non-whitened inputs can be whitened through a linear transformation (see Eq. 2.8), second-order invariant learning rules will not be sensitive to second-order correlations in the data. As we have seen in the previous chapter, nonlinear Hebbian learning rules can learn receptive fields from whitened inputs, under constraints on norm of the weights. As whitening and norm constraints are implicitly handled by second-order invariant learning rules, they should learn receptive fields directly from non-whitened data.

In Figure 3.2a, we show that second-order plasticity learns localized receptive fields from non-whitened natural images. In comparison, a learning rule with multiplicative synaptic depression

$$\Delta \mathbf{w} \propto \mathbf{x}\, y^2 - \mathbf{w}\, y^2 \tag{3.5}$$

does not enable second-order invariance, leading receptive fields determined by the principal components of the data (Figure 3.2b).

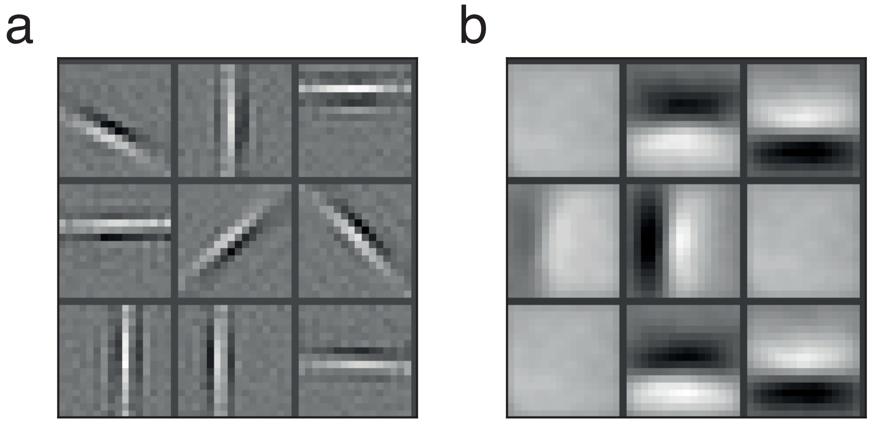

Figure 3.2 – **Receptive field formation from non-whitened inputs.** (**a**) Nine independent trials of second-order invariant learning ($h(y) = y^2$) for non-whitened natural images lead to localized oriented receptive fields. (**b**) Nine independent trials for nonlinear Hebbian learning with multiplicative LTD (Eq. 3.5) lead to principal components.

3.3 Discussion

3.3.1 Relation to previous studies

Hebbian models such as BCM and Oja's rules are decades old, and a huge number of studies have investigated their functional properties, in terms of stability, feature selectivity and receptive field development (Cooper et al., 2004). These studies can now be unified in a single theoretical framework. For instance, the BCM model has been observed to learn receptive fields even in the absence of input pre-processing (Law and Cooper, 1994). This empirical finding is now explained by the presented theory of second-order invariance.

The BCM model has already been previously linked to projection pursuit and higher-order statistical learning (Intrator and Cooper, 1992). However, these analyses have missed that the objective function of BCM is second order invariant, being equivalent to stable normalized projection pursuit. These theoretical ambiguities are clear in BCM studies that have empirically compared receptive fields developed by BCM and normalized projection pursuit (Blais et al., 1998). With our theory in hand, their empirical results can be predicted a priori.

The functional difference between Oja's rule heterosynaptic depression and BCM's meta-plasticity has also been unclear (Cooper et al., 2004). Our analysis shows that meta-plasticity is a requirement for second-order invariance, while Oja's heterosynaptic depression has only multiplicative scaling effects, providing stability, but no additional selectivity properties.

We believe that our findings provide a systematic method for the analysis and development of Hebbian plasticity models.

3.3.2 Robustness to network heterogeneities

An interesting corollary of our analysis is that independent linear scaling of pre-synaptic inputs can be compensated by synaptic plasticity. For instance, pre-synaptic neurons may have diverse average firing rates. Since average firing rates are linear properties, synaptic weights will compensate for these differences, e.g. by down-regulating synapses that receive inputs from highly active neurons.

Another source of heterogeneity across synaptic inputs is due to dendritic attenuation of EPSPs. EPSPs originating from synapses at distal dendrites have lower impact at the somatic membrane potential. However, it has been observed that distal synapses are relatively up-regulated, and have in general somatic effects at the same order of magnitude than proximal connections (Magee and Cook, 2000). Second-order invariant plasticity gives a simple explanation for these findings.

3.3.3 Functional interpretation of cortical plasticity

Our analytical derivations provide a novel perspective for the functional interpretation of synaptic potentiation, depression and homeostasis. We have shown in the previous chapter that a nonlinear Hebbian term is capable of universal feature learning, which is implemented by nonlinear LTP. While the specific form the nonlinear function was left open, we speculate that cortical plasticity may implement the simplest function that would satisfy the functional requirements. Since the third moment is the smallest higher-order moment, its implementation could be motivated by this simplicity argument.

We have shown here that LTD also has a clear functional interpretation, providing invariance to second-order statistical properties of the input. We may suggest an intuition for how this property comes about. LTD is modeled as a linear anti-Hebbian factor. While linear Hebbian learning implements variance maximization (Oja, 1989), linear anti-Hebbian learning will minimize variance. Thus the LTD term cancels the tendency of the LTP factor to follow directions of higher variance, making it only selective to higher order statistics.

The homeostatic factor has also a precise functional interpretation. It must balance LTP and LTD, so that LTD cancels the correct amount of variance dependency present in the LTP term. This implies that if more mechanisms are added to the plasticity rule, for instance heterosynaptic plasticity, as in Oja's learning rule, $-\mathbf{w}\, y^2$, there will be an imbalance between the LTP and LTD factors, and second-order invariance will be lost. It shows that meta-plasticity, as in BCM models, has functional properties absent in alternative homeostatic mechanisms, such as multiplicative scaling (Oja, 1989; Turrigiano, 2011).

Our results also apply to phenomenological models of spike-timing dependent plasticity. For instance, under the Poisson firing assumption, the triplet STDP model reduces to a quadratic LTP factor, and a linear LTD factor (Pfister and Gerstner, 2006), consistent with the models we have analyzed. Our theory constrains synaptic depression to be linear on the pre- and post-synaptic activities, suggesting that pairing experiments under Poisson firing times of pre- and post-synaptic neurons would be valuable to investigate to what extent these properties hold in biology (Froemke and Dan, 2002).

3.4 Methods

3.4.1 Linear invariance of normalized projection pursuit

Consider any optimization objective of the form

$$\mathbf{w}^* = \text{argmax}_{\mathbf{w}} \left\langle F\left(\frac{\mathbf{w}^T\mathbf{x}}{\sigma}\right)\right\rangle \tag{3.6}$$

with $\sigma = \sqrt{\left\langle (\mathbf{w}^T\mathbf{x})^2 \right\rangle}$. Let $\mathbf{M}$ be a whitening matrix for $\mathbf{x}$, so that $\tilde{\mathbf{x}} = \mathbf{M}\mathbf{x} \implies \left\langle \tilde{\mathbf{x}}\tilde{\mathbf{x}}^T \right\rangle = I$ (c.f. Eq. 2.8). Using that $\mathbf{x} = \mathbf{M}^{-\mathbf{1}}\tilde{\mathbf{x}}$ and defining $\tilde{\mathbf{w}} = \mathbf{M}^{-T}\mathbf{w}$, we have

$$\begin{aligned}
\left\langle F\left(\frac{\mathbf{w}^T\mathbf{x}}{\sigma}\right)\right\rangle &= \left\langle F\left(\frac{\mathbf{w}^T\mathbf{M}^{-\mathbf{1}}\tilde{\mathbf{x}}}{\sqrt{\left\langle (\mathbf{w}^T\mathbf{M}^{-\mathbf{1}}\tilde{\mathbf{x}})^2 \right\rangle}}\right)\right\rangle && (3.7)\\
&= \left\langle F\left(\frac{\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}}{\sqrt{\left\langle (\tilde{\mathbf{w}}^T\tilde{\mathbf{x}})^2 \right\rangle}}\right)\right\rangle && (3.8)\\
&= \left\langle F\left(\frac{\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}}{|\tilde{\mathbf{w}}|}\right)\right\rangle && (3.9)\\
&= \left\langle F\left(\frac{\tilde{\mathbf{w}}^T}{|\tilde{\mathbf{w}}|}\tilde{\mathbf{x}}\right)\right\rangle && (3.10)
\end{aligned}$$

where we used that $\left\langle (\tilde{\mathbf{w}}^T\tilde{\mathbf{x}})^2 \right\rangle = \left\langle \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T\tilde{\mathbf{w}} \right\rangle = \tilde{\mathbf{w}}^T \left\langle \tilde{\mathbf{x}}\tilde{\mathbf{x}}^T \right\rangle \tilde{\mathbf{w}} = \tilde{\mathbf{w}}^T\tilde{\mathbf{w}} = |\tilde{\mathbf{w}}|^2$. Thus the normalized projection pursuit can be mapped to a standard projection pursuit, with normalized weights, over the whitened inputs $\tilde{\mathbf{x}}$,

$$\tilde{\mathbf{w}}^* = \text{argmax}_{\tilde{\mathbf{w}},|\tilde{\mathbf{w}}|=1} \left\langle F\left(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}\right)\right\rangle \tag{3.11}$$

with an optimum in the original input space given by $\mathbf{w}^* = \mathbf{M}^T\tilde{\mathbf{w}}^*$.

Analogously, given any linear transformation of the input, $\mathbf{x}' = \mathbf{R}\mathbf{x}$, for an invertible matrix $\mathbf{R}$, we may map the normalized projection pursuit to the whitened projection pursuit of Eq. 3.11, with the optima given by $\mathbf{w}'^* = \mathbf{R}^{-T}\mathbf{M}^T\tilde{\mathbf{w}}^*$.

3.4.2 Derivation of second-order invariant learning rule

We consider the normalized projection pursuit objective,

$$\mathbf{w}^* = \text{argmax}_{\mathbf{w}} \left\langle \left(\frac{y}{\sigma_y}\right)^3 \right\rangle \tag{3.12}$$

Proceeding with gradient ascent on $\mathbf{w}$, and using that $\frac{\partial \sigma_y}{\partial \mathbf{w}} = \frac{\partial \sqrt{\langle y^2 \rangle}}{\partial \mathbf{w}} = \langle y \mathbf{x}_+ \rangle / \sigma_y$,

$$\frac{\partial \langle F \rangle}{\partial \mathbf{w}} = \langle \frac{\partial \left(\frac{y}{\sigma_y}\right)^3}{\partial \mathbf{w}} \rangle \tag{3.13}$$

$$= 3\langle \left(\frac{y}{\sigma_y}\right)^2 \left(\sigma_y^{-1} \frac{\partial y}{\partial \mathbf{w}} + y \frac{\partial \sigma_y^{-1}}{\partial \mathbf{w}}\right) \rangle \tag{3.14}$$

$$= \frac{3}{\sigma_y^2} \langle y^2 \left(\frac{\mathbf{x}_+}{\sigma_y} - \frac{y}{\sigma_y^2} \frac{\partial \sigma_y}{\partial \mathbf{w}}\right) \rangle \tag{3.15}$$

$$= \frac{3}{\sigma_y^2} \langle \left(\frac{\mathbf{x}_+ y^2}{\sigma_y} - \frac{y^3}{\sigma_y^3} \langle \mathbf{x}_+ y \rangle \right) \rangle \tag{3.16}$$

$$= \frac{3}{\sigma_y^3} \langle \left(\mathbf{x}_+ y^2 - \frac{y^3}{\langle y^2 \rangle} \langle \mathbf{x}_+ y \rangle \right) \rangle \tag{3.17}$$

$$= \frac{3}{\sigma_y^3} \left(\langle \mathbf{x}_+ y^2 \rangle - \frac{\langle y^3 \rangle}{\langle y^2 \rangle} \langle \mathbf{x}_+ y \rangle \right) \tag{3.18}$$

where we defined $\mathbf{x}_+$ as the input for samples in which $y > 0$. But since for samples in which $y \leq 0$ the update is zero, we may replace $\mathbf{x}_+$ by $\mathbf{x}$.

We may consider a separation of time scales, and assume that the estimation of σ_y and $\frac{\langle y^3 \rangle}{\langle y^2 \rangle}$ is performed at a faster time scale than the other terms, which allows us to consider them as constants. We may now transform this offline learning rule into a stochastic gradient ascent version, by removing the estimation over the whole dataset,

$$\Delta \mathbf{w} \propto \mathbf{x}\, y^2 - h^*(y)\, \mathbf{x}\, y \tag{3.19}$$

where $h^*(y) = \frac{\langle y^3 \rangle}{\langle y^2 \rangle}$ is referred to as the balancing homeostatic factor.

3.4.3 Alternative homeostatic factors maintain second-order invariance

We calculate the homeostatic factor after the norm has converged to a stable value. Under this assumption the gradient in the direction of the weights $\mathbf{w}$ is zero:

$$\langle \mathbf{w}^T \Delta \mathbf{w} \rangle \propto \langle y^3 \rangle - h(y) \langle y^2 \rangle = 0 \implies h(y) = \langle y^3 \rangle / \langle y^2 \rangle = h^*(y) \tag{3.20}$$

We can also calculate analytically the value of $h(y)$ at the points where the norm is stable. For the example of $h(y) = \langle y^2 \rangle$, we have

$$\begin{aligned} h(y) = \langle y^3 \rangle / \langle y^2 \rangle &\iff \langle y^2 \rangle = \langle y^3 \rangle / \langle y^2 \rangle && (3.21) \\ &\iff |\mathbf{w}|^2 \langle x_w^2 \rangle = |\mathbf{w}| \langle x_w^3 \rangle / \langle x_w^2 \rangle && (3.22) \\ &\iff |\mathbf{w}| = \langle x_w^3 \rangle / \langle x_w^2 \rangle^2 && (3.23) \end{aligned}$$

where $x_w = (\mathbf{w}^T \mathbf{x})_+ / |\mathbf{w}|$ is the rectified projection of the input $\mathbf{x}$ on the normalized direction $\mathbf{w}/|\mathbf{w}|$.

4 Estimation theory for stochastic optimization and adaptive learning rates

In stochastic gradient descent (SGD), samples of the gradient of a loss function are obtained one data sample at a time, and modulated by a learning rate factor. Gradient samples are variable and it is unclear how this variability may influence parameter learning dynamics. We show that the signal-to-noise ratio of gradient samples determines an optimal batch size for which the gradient is reliably estimated. We show that this optimal batch size can be implemented implicitly in SGD through a novel adaptive learning rate algorithm, which maintains the magnitude of parameter change stable across parameters and across time, independent of the gradient statistics. It also provides a bound on learning time for SGD, revealing a bottleneck in learning dynamics complementary to constraints on the optimization surface. These results provide a unifying framework for understanding adaptive learning rate models and the learning dynamics in stochastic gradient descent.

4.1 Introduction

Learning in complex systems such as neural networks involves the optimization of a large number of parameters. Since analytical solutions are not possible, gradient descent is the most common optimization method. In offline gradient descent the full dataset is used to estimate the gradient at each iteration of the algorithm, but in modern applications of machine learning the datasets required for learning can be immense. In such situations, one typically implements stochastic gradient descent, in which each parameter update uses only one, or a small number, of data samples (Bottou, 2010; Bengio et al., 2013; Mnih et al., 2013). In the previous chapters, we've shown that cortical synaptic plasticity can also be interpreted as performing stochastic gradient descent (SGD), by updating synaptic weights continually based on the noisy neural activity, and only on average does it optimize a particular goal (Frémaux et al., 2010; Pfister et al., 2006; Seung, 2003).

A fundamental concern in SGD is the appropriate size of the learning rate factor which modulates the update step for each gradient sample. Since each gradient sample is highly variable, big update step sizes lead to random excursions in the parameter space that will not follow the true gradient direction. On the other hand, small step sizes slow down learning. Further complications arise because the gradient sample statistics are not uniform in a model, but may change in time or across parameters. Such statistical heterogeneities imply that a fixed learning rate for all parameters, at all times, will not be optimal.

This problem has led to the formulation of adaptive learning rates, that modulate the update step size depending on the statistics of the gradient samples (Duchi et al., 2011; Schaul et al., 2012; Kingma and Ba, 2014; Dauphin et al., 2015; Duchi et al., 2011). Here we present a theory that explains the learning properties of SGD algorithms by interpreting stochastic optimization as efficient estimation of the gradient from a limited number of samples.

Based on this estimation theory, we develop a novel adaptive learning rate that implicitly implements the optimal batch size for robust gradient estimation. It makes the effective parameter change independent of the heterogeneities in the gradient sample statistics, across parameters and across time. We show that our model outperforms previous adaptive algorithms in a convex optimization problem, and that our theory provides a unifying framework for understanding adaptive learning rates in stochastic optimization. In particular, it suggests how synaptic plasticity may implement adaptive learning.

Estimation theory also reveals an appropriate way of analyzing how parameters evolve in SGD. We show that learning time is modulated by the surface of the gradient sampling statistics, in contrast with the true gradient surface, commonly studied in optimization theory (Hiriart-Urruty and Lemarechal, 2013; Dauphin et al., 2014). Our results allow us to predict the learning dynamics and the convergence times of SGD, providing a general theory for stochastic optimization.

4.2 Results

4.2.1 Sampling size for a reliable gradient estimation

We consider the goal of finding parameters Θ that minimize function $F_\Theta(x)$ over a dataset $X = (x_1, \ldots, x_N)$,

$$\min_\Theta \langle F_\Theta(x) \rangle_x = \min_\Theta \frac{1}{N} \sum_{x \in X} F_\Theta(x) \,. \tag{4.1}$$

In gradient descent we update each parameter $\theta \in \Theta$ iteratively, $\theta_{k+1} = \theta_k - \eta \; g_k$, where $g_k = \left\langle \frac{\partial F_{\theta_k}(x)}{\partial \theta} \right\rangle_x$ and k denotes the time step. The learning rate η modulates the step size of the parameter updates. Choosing the same learning rate η for all parameter updates is often suboptimal, which motivates the search for methods that determine the optimal individual learning rates[1].

[1] If each parameter update has its own learning rate, one speaks also of hill descent or gradient descent under a non-Euclidian metric.

A large portion of previous theoretical efforts aim to adapt the learning rate to an optimal step size based on the geometry of the optimization surface, such as curvature (Hiriart-Urruty and Lemarechal, 2013; Dauphin et al., 2014; Bottou, 2010). Here we would like to separate this problem from another fundamental obstacle in online optimization, which is how to estimate the gradient g_k from noisy samples.

In stochastic gradient descent (SGD), instead of calculating the gradient using the whole dataset, one performs updates sequentially for each data sample x_t, using the gradient sample $g_t = \frac{\partial F_\Theta(x_t)}{\partial\theta}$. Since g_t can be highly variable, many samples may be needed for a reliable estimate of the true gradient. Thus we ask the following question: how many data samples are sufficient for a reliable estimation of the true gradient g_k?

For the moment, we disregard the question of step size. Instead we only want to know if a specific parameter θ should be increased or decreased to minimize $\langle F\rangle$, which depends on the sign of the gradient, $g_k = \langle\frac{\partial F_{\theta_k}}{\partial\theta}\rangle$. Once the gradient direction is known, we may adjust the step size to a desired update magnitude. Here we consider fixed update magnitudes η_0 and assume that each parameter update leads only to small relative change in the true gradient:

$$\left|\left\langle\frac{\partial F_\theta}{\partial\theta}\right\rangle_x - \left\langle\frac{\partial F_{\theta\pm\eta_0}}{\partial\theta}\right\rangle_x\right| \ll \left|\left\langle\frac{\partial F_\theta}{\partial\theta}\right\rangle_x\right| \tag{4.2}$$

Before we turn to classical online stochastic gradient descent, let us study the case of updates after having collected batches of B data samples with a fixed parameter setting Θ. A gradient sample at time t is an unbiased estimate of the gradient, $\mu = \langle g_t\rangle$. We assume that gradient samples g_t have high variability, $\sigma \gg \mu$, where σ is the standard deviation of the distribution of gradient samples. In this case, a large batch of data is needed for a reliable estimation of the gradient. For a batch size B, the gradient estimator $\tilde{\mu}$ is

$$\tilde{\mu} = \frac{1}{B}\sum_{t=1}^{B} g_t \tag{4.3}$$

where B is the number of samples we take for fixed parameters Θ.

For independent samples and a large batch size B, the central limit theorem suggests that the estimate is approximately normally distributed

$$\tilde{\mu} \sim N(\mu, \sigma/\sqrt{B}) \tag{4.4}$$

Intuitively, we expect $\tilde{\mu}$ to be a reliable estimate of μ when the mean of the distribution becomes larger than the variability, $\mu \geq \sigma/\sqrt{B}$, implying a constraint on the batch size

$$B \geq \frac{\sigma^2}{\mu^2} = B^* \tag{4.5}$$

where we have defined the critical batch size $B^* = \frac{\sigma^2}{\mu^2}$. In other words, if we take B^* data samples the estimate of the gradient is $\tilde{\mu} \sim |\mu| \cdot N(\pm 1, 1)$.

We may formalize this result as a decision problem. If we impose a probability α of estimating the correct gradient sign at each batch iteration, then we need a batch size

$$B = c_\alpha \, \frac{\sigma^2}{\mu^2} \tag{4.6}$$

where $c_\alpha = \frac{z^2_{\alpha/2}}{4}$ is a constant dependent on the confidence level α, with $z_{\alpha/2}$ the $100 * (1 - \alpha/2)$ percentile of the normal distribution. For instance, for twice the critical batch size, $c_\alpha = 2$, the confidence level is $\alpha = 0.97$.

We can also relate this result to the information content in the gradient sample. If we consider a normal prior for the gradient, $\mu^* \sim N(0, |\mu|)$, the mutual information between B gradient samples and the gradient sign is

$$I(g_{1..B}; \mu^*) \propto B \, \frac{\mu^2}{\sigma^2} \tag{4.7}$$

leading again to the relation $B = c_\alpha \, \frac{\sigma^2}{\mu^2}$, with the constant c_α dependent on level of information desired.

We conclude that the number of samples required for a reliable estimate of the gradient's sign is inversely proportional to the signal-to-noise ratio (SNR) of the gradient samples. For batch sizes smaller than the critical value, $B \ll B^*$, the sign estimation is highly variable, while a larger batch size, $B \gg B^*$, will have diminishing returns on the estimates reliability.

4.2.2 Drift-diffusion model for stochastic gradient descent

In the previous section, we assumed that we take several data samples, before we update the parameters. We now would like to extend our estimation analysis to online learning, implemented through stochastic gradient descent (SGD). In SGD the gradient samples are processed continuously, with updates performed at each sample x_t,

$$\theta_{t+1} = \theta_t + \eta_t \, g_t \tag{4.8}$$

This process can be interpreted as a drift-diffusion process. Maintaining the assumption of high noise and small updates, we approximate the model as a Gaussian process, with update steps given by $\eta_t \, g_t \sim \eta_t \, N(\mu_t, \, \sigma_t)$.

In the previous section, we assumed that parameters remained fixed while we draw B gradient samples. Note that B can be significantly smaller than the total number of data available, and we have still a reliable gradient estimate.

Our objective is to retain the property of a reliable estimate of the gradient as defined in the previous section. To do so, we use a separation of time scales. Because of our assumption (4.2), we can find a time frame T over which parameters do not change significantly, so that we can assume stationary statistics for g_t over T. We calculate the distribution of the update after $T = B$ samples, where update steps are of size $\eta^* < \eta_0$,

$$\Delta^B \theta = \sum_{t=1}^{B} \eta \, g_t \sim N(\eta \, B \, \mu, \, \eta \, \sqrt{B} \, \sigma) \,. \tag{4.9}$$

By searching for values of η^* and B^*, so that the magnitude is $\eta^* \; B^* \; |\mu| = \eta_0$, and the variability $\eta^* \; \sqrt{B^*} \; \sigma = \eta_0$ is of the same order, we arrive at

$$\eta^* = \eta_0 \, \frac{|\mu|}{\sigma^2} \,. \tag{4.10}$$

After $B^* = \frac{\sigma^2}{|\mu|^2}$ update steps the total change is

$$\Delta^{B^*} \theta = \sum_{t=1}^{B^*} \eta^* \; g_t \sim \eta_0 \; N(\pm 1, 1) \,. \tag{4.11}$$

Thus this online rule behaves like the batch update. We refer to SGD with 4.10 as Sampa.

In Fig. 4.1, we illustrate the effect of the learning rate η_0 on the gradient estimation and effective step size for a one dimensional SGD example. Sampa ensures after B^* gradient samples, the parameter will have changed on the scale of η_0 (Fig. 4.1b). Larger learning rates than indicated by Sampa imply unreliable changes of direction on the order of magnitude η_0 (Fig. 4.1a). Smaller learning rates will imply smaller changes after B^* samples, and will demand more samples to reach the η_0 scale (Fig. 4.1c).

4.2.3 Sampa: Sampling-based adaptive learning rate

The benefits of the proposed adaptive learning rate come at the cost of requiring robust estimates of the first- and second-order statistics of the gradient samples, $|\mu|^*$ and σ^{2^*}, at each time step. We can implement both estimates as moving averages

$$\mu^*_{t+1} = (1 - 1/\tau_\mu) \; \mu^*_t + g_t / \tau_\mu \tag{4.12}$$

$$\sigma^{2^*}_{t+1} = (1 - 1/\tau_\sigma) \; \sigma^{2^*}_t + g_t^2 / \tau_\sigma \tag{4.13}$$

Again we assumed $\sigma \gg \mu$, in which case the variance of g_t is well approximated by its second-order moment, i.e. we do not subtract the mean.

Our previous estimation analysis tells us that the time scale for the robust estimation of $|\mu|$

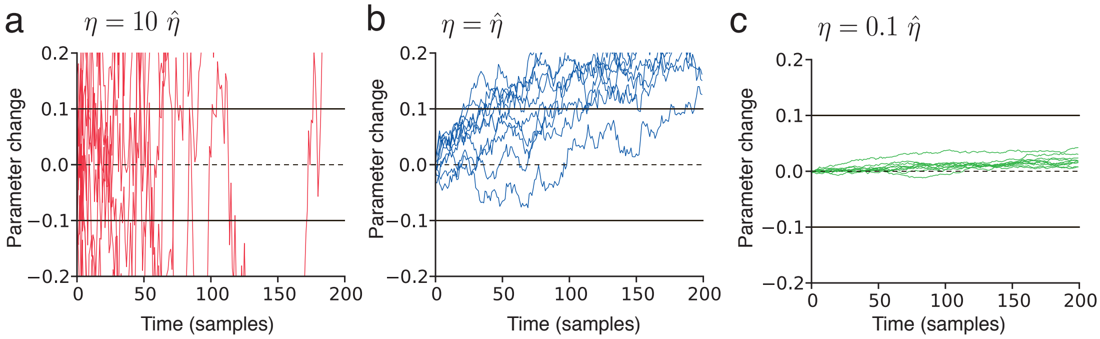

Figure 4.1 – **Reliable gradient estimation with adaptive SGD.** Gradient samples were generated with $\mu = 0.5$ and $\sigma = 5.0$, for which the critical estimation batch size is $B^* = 100$ samples. We assume a desired effective step size $\eta_0 = 0.1$, for which we have a Sampa learning rate of $\eta^* = 0.002$. (**a**) For learning rate larger than Sampa, $\eta = 10\,\eta^*$, the update is highly unreliable in the scale of η_0. (**b**) For the Sampa learning rate, $\eta = \eta^*$, after the critical sample size $B^* = 100$, the parameter will have moved on the order of η_0, and with high probability in the correct direction. (**c**) For smaller learning rates, $\eta = 0.1\,\eta^*$, after B^* samples the correct direction has already been estimated, but the parameter has been updated by only $0.1\eta_0$, and will require ten times more samples to update on the order of η_0.

should be B^*, so we set

$$\tau_\mu = B^* = \sigma^2/\mu^2 \tag{4.14}$$

Because of Eqs. 4.2 and 4.9, the magnitude of μ does not significantly change during B^* samples (that is, on the order of η_0 change in the parameter). The estimation of σ^2 is in general not as sensitive, since $\sigma \gg \mu$, and we choose a timescale of $\tau_\sigma = 1000$ samples.

For a sensible initialization, we propose to use for both estimates a batch of T_0 gradient samples with the initial parameters θ_0

$$\mu_0^* = \frac{1}{T_0}\sum_{t=1}^{T_0} \frac{\partial F_{\theta_0}}{\partial \theta}(x_t) \tag{4.15}$$

$$\sigma_0^{2^*} = \frac{1}{T_0}\sum_{t=1}^{T_0} \left(\frac{\partial F_{\theta_0}}{\partial \theta}(x_t)\right)^2 \tag{4.16}$$

where we have used $T_0 = 10000$ in this study. The implementation of Sampa is described in Algorithm 1.

The effective step size η_0 in Sampa has a conveniently explicit meaning: it defines the precision of the parameter search. It implies (i) that the algorithm assumes that the optimization surface does not change widely on the scale of η_0, and (ii) after convergence the parameters will fluctuate around the optimum with a variability on the order of η_0. In many cases, parameters are constrained within a certain range, e.g. $0 \le |\theta| \le \theta_{max}$, from which one may establish a desired precision, e.g., $\eta_0 = 0.001\,\theta_{max}$.

We choose a linear regression problem to analyze the behavior of Sampa in a convex setting. We generate data samples (x^t, y^t) from $y^t = \hat{w}_1 x_1^t + \hat{w}_2 x_2^t + \epsilon$, with $x_1 \sim N(0,1)$, $x_2 \sim N(0,2)$ and $\epsilon \sim N(0,3)$. The objective is to find the optimal parameters $(\hat{w}_1, \hat{w}_2)$ that minimize the squared error, $\langle F(w_1, w_2)\rangle = \langle (y - w_1 x_1 + w_2 x_2)^2 \rangle$. We will use this example throughout our study, with true values $\hat{w}_1 = 1$, $\hat{w}_2 = -1$ and initial parameters $w_1^0 = -3$, $w_2^0 = -4$.

In Fig. 4.2 we compare the robustness of its implementation to its oracle version, which uses the true values of $|\mu|$ and σ^2 at each time step, instead of the estimates $|\mu|^*$ and σ^{2^*}.

Both versions follow similar learning dynamics (Fig. 4.2a), with a delay for Sampa compared to its oracle version (Fig. 4.2b). This delay is explained by the low-pass filter in the estimates μ^* and σ^{2^*} in Eqs. 4.15 and 4.16 (Fig. 4.2c-d).

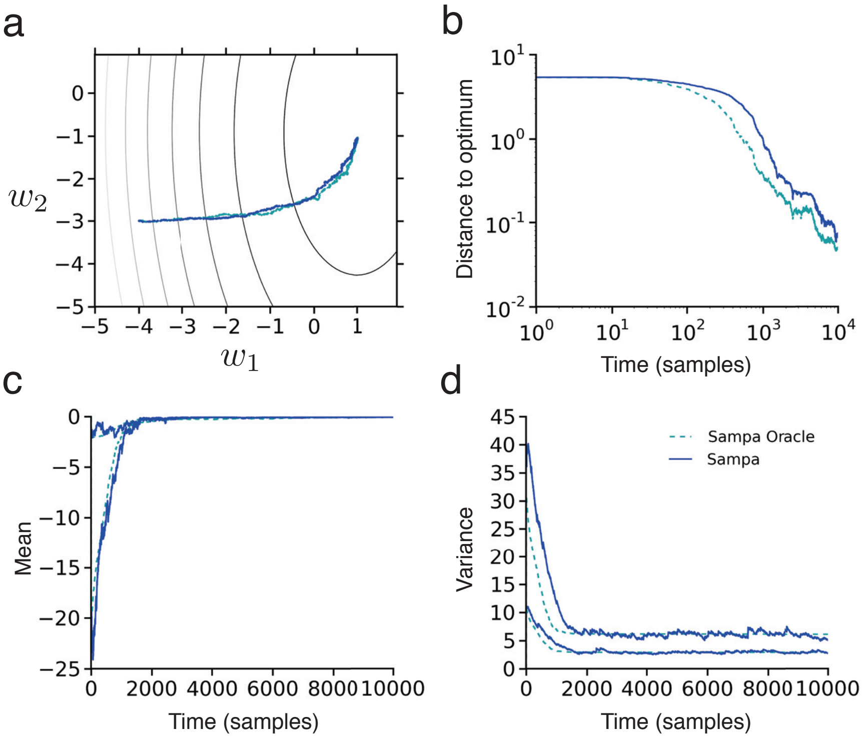

Figure 4.2 – **Implementation of the adaptive method Sampa in a convex 2D problem.** (**a**) Background contours indicate parameter values with constant optimization error. Sampa (dark blue) and its oracle version (light blue) have similar parameter trajectories, converging to the optimum ($\eta_0 = 0.03$). (**b**) Euclidean distance to the optimum in time, $|\mathbf{w}_t - \mathbf{w}_{\text{opt}}|_2$, for Sampa and its oracle version. Sampa has similar learning dynamics to its oracle, albeit with a delay. (c-d) The Sampa estimates of μ_t^* (**c**) and $\sigma_t^{2^*}$ (**d**) for both parameters follow the true estimates μ_t with a delay, explaining the delay in the learning dynamics.

Algorithm 1 Sampling-based adaptive learning rate: Sampa

```
procedure MINIMIZE $\langle F(x) \rangle_x$
    Initialize $\theta$
    Choose $\eta_0$, $\tau_\sigma$, $T_0$, $\epsilon$.
    $\mu \leftarrow \frac{1}{T_0} \sum_{t=1}^{T_0} \partial F_\theta(x_t)$
    $\sigma^2 \leftarrow \frac{1}{T_0} \sum_{t=1}^{T_0} \partial F_\theta(x_t)^2$
    for each data sample $x_t$, do
        for each parameter $\theta$, do
            $g \leftarrow \partial F_\theta(x_t)$
            $\tau_\mu \leftarrow \sigma^2 / \mu^2$
            $\eta \leftarrow \eta_0 \, |\mu| / \sigma^2$
            $\mu \leftarrow (1 - 1/\tau_\mu)\, \mu + g/\tau_\mu$
            $\sigma^2 \leftarrow (1 - 1/\tau_\sigma)\, \sigma^2 + g^2/\tau_\sigma$
            $\theta \leftarrow \theta - \eta\, g$
```

4.2.4 Comparison between Sampa and simpler adaptive models

We compare Sampa ($\eta_{Sampa} = \eta_0 \, |\mu|/\sigma^2$ with two simpler models, standard SGD, with fixed learning rate ($\eta_{SGD} = \eta_0$), and Rmsprop ($\eta_{Rms} = \eta/\sigma$). Rmsprop is a commonly used adaptive model which implements variance normalization (Tieleman and Hinton, 2012), dividing the learning rate by the standard deviation of the update steps.

In Fig. 4.3a we see that Sampa scales down the step size as the parameters approach the optimum. This behavior follows from the fact that near the optimum the gradient μ is small, while the gradient variability stays at the same order of magnitude as before. In this case a smaller step size is needed for an effective estimation of the gradient μ. Since the Sampa update step is proportional to $|\mu|/\sigma^2$ it does the desired rescaling automatically. Smaller η_0 leads to slower convergence, but higher precision of the final parameter values (Fig. 4.3b). As predicted by our theory, the variability after convergence is on the scale of η_0, since it determines the effective resolution of the parameter search.

Standard SGD exhibits, for similar initial step sizes, much more variability after convergence (Fig. 4.3c,d). This follows from the lack of adaptation to the signal-to-noise ratio of the gradient, so that near the optimum the learning rate is not decreased.

Rmsprop is invariant to the variance of the gradient, but not to the signal-to-noise ratio. Thus, it does not scale the step size near the optimum, showing high variability at convergence, similar to standard SGD.

The observed jitter in the parameter trajectory can be estimated by our theory for each adaptive model η. For B^* samples, the effective parameter change $\Delta\theta^*$ is given by

$$\Delta\theta^* = \pm \frac{\eta^t}{|\mu^t|/{\sigma^t}^2} \tag{4.17}$$

It explains why standard SGD and Rmsprop have high variability near convergence, where the gradient is small, since their parameter change is modulated by $1/|\mu^t|$.

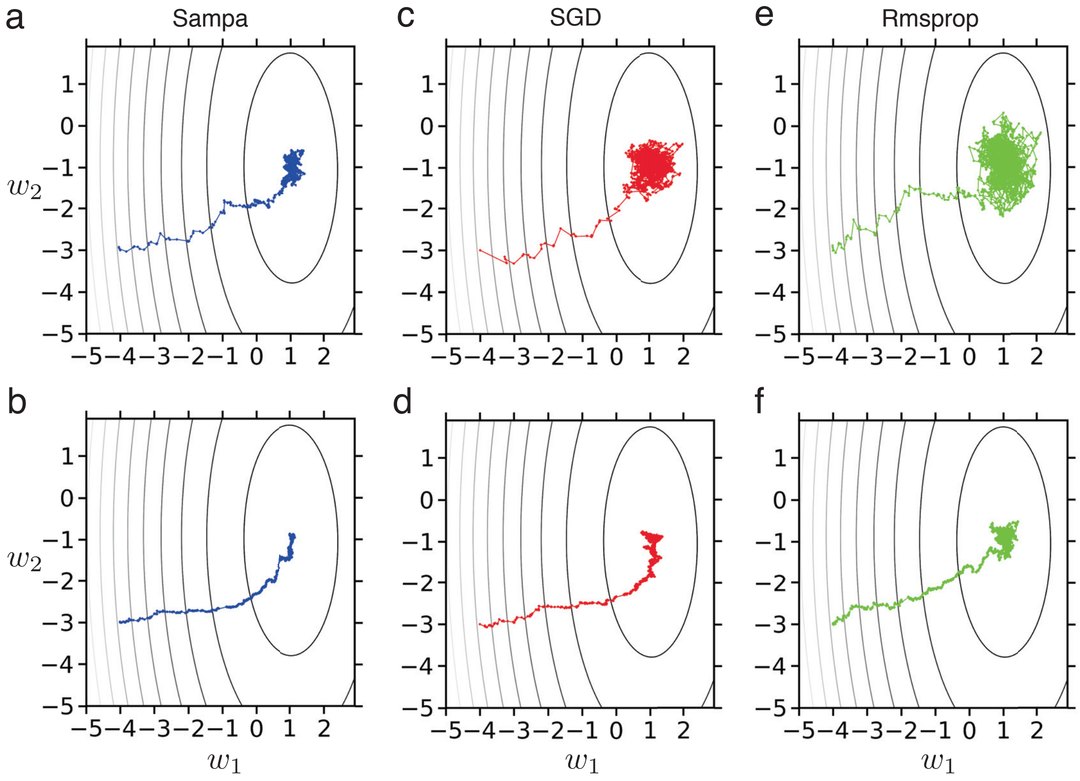

Figure 4.3 – **Comparison of the variability of Sampa, SGD and Rmsprop near convergence.** (**a,c,e**) For similar initial step sizes, Sampa shows less variability after convergence than SGD and Rmsprop. It follows from Sampa's adaptation to the small gradients near the optima. (**b,d,f**) Smaller initial step sizes leads to proportionally less variability at the optima for all models, with Sampa showing less variability than SGD and Rmsprop.

The trade off between convergence speed and precision is not particular to SGD, but also affects offline gradient descent. What is special about Sampa is that it can maintain a constant parameter change $\Delta\theta^*$ throughout the simulation, as it is possible in offline gradient descent. The same is not achieved by standard SGD or Rmsprop, where the effective parameter change $\Delta\theta^*$ is sensitive to the signal-to-noise ratio of the gradient, varying across time.

For a comparison of convergence between the three models, we measure the distance to the optimum for each model after a T samples. We define the distance d_T as the value for which, at time T, the parameters are within that distance in at least 95% of the trials. To allow a fair comparison, we use for each T the best meta-parameter η_0 for a given model. Fig. 4.4 shows that Sampa has a better performance than both SGD and Rmsprop over the range of learning durations inspected.

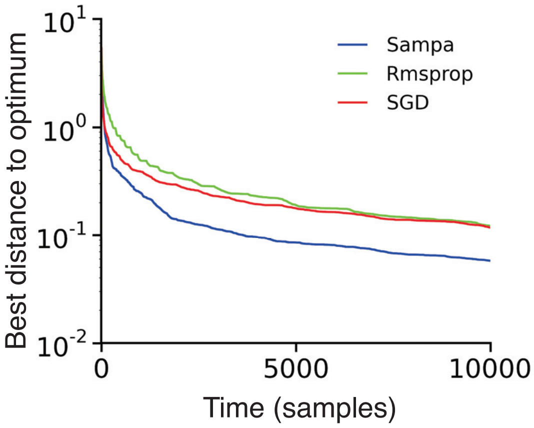

Figure 4.4 – **Comparison of convergence speed between Sampa, SGD and Rmsprop.** Sampa shows faster convergence than SGD and Rmsprop.

4.2.5 Estimation theory for momentum and mini-batches

Estimation theory provides an interesting interpretation of two prevalent techniques in gradient optimization: momentum and mini-batches. With momentum, the gradient samples are low-pass filtered

$$\bar{g}_{t+1} = (1 - 1/\tau_{mom})\ \bar{g}_t + g_t/\tau_{mom} \tag{4.18}$$

$$\Delta\theta_{t+1} = \eta\ \bar{g}_{t+1} \tag{4.19}$$

The implementation of momentum has been associated with the avoidance of local minima or parameter oscillations, as may occur near narrow valleys in the optimization surface (Hertz et al., 1991). These are valid arguments in favor of using momentum in the batch mode.

However, we argue that, in SGD, momentum is well justified as smoothing of the rather noisy sampling of the gradient. Our theory gives a foundation for the time scale τ_{mom}, which normally is chosen heuristically. As B^* samples are needed to estimate the gradient, the variability in the parameter trajectory within this time frame may be viewed as sampling noise. Thus by setting the adaptive momentum time scale to

$$\tau_{mom} = B^* = \sigma^2/\mu^2 \tag{4.20}$$

one will estimate the gradient with a smooth and well calibrated moving average. It may be implemented by replacing line 13 in Algorithm 1 by

$$\theta \leftarrow \theta - \eta\ \mu. \tag{4.21}$$

Since the stochastic fluctuations in Sampa are at scales smaller than η_0, it may be that they do not interfere, and it remains an empirical question if the adaptive momentum adds to the stability of Sampa in real-world applications.

Mini-batches is another common form of gradient averaging used in SGD, in which updates use b gradient samples, at each iteration k,

$$\bar{g}_k = \frac{1}{b} \sum_{t=t_k}^{t_k+b} g_t \tag{4.22}$$

$$\Delta\theta_k = \eta\ \bar{g}_k \tag{4.23}$$

Not only it attenuates gradient sampling noise, but it is a crucial property for large-scale applications, since the simultaneous processing of many data points allows for vectorial and parallel operations in the simulation.

4.2.6 Convergence time of stochastic optimization

Our gradient estimation theory also has important implications for the understanding of the learning dynamics of SGD. It provides a tool for estimating the path that parameters will follow during optimization. For each parameter θ, the velocity $v_\theta = \theta_{t+1} - \theta_t$ is

$$v_\theta = \eta_{Sampa}\ \mu_\theta = \eta_0 \ \mathrm{sign}(\mu_\theta)\ \frac{\mu_\theta^2}{\sigma_\theta^2} \tag{4.24}$$

The speed is therefore constrained by the sample size required for an effective update step.

In Fig. 4.6a, we plot the velocity field arising from Eq. 4.24 for our 2D example, and illustrate how a trial implementing Sampa follows roughly the predicted path. Thus for SGD it is not the true gradient surface that guides the parameter search, but the geometry of the gradient's signal-to-noise ratio surface.

Importantly, from the velocity we can estimate the learning time. Since B^* samples are required at each step size η_0, the learning time T to go from an initial parameter position θ_0 to another, θ_*, will given by

$$T = \frac{1}{\eta_0} \int_{\theta_0}^{\theta_*} \frac{\sigma_\theta^2}{\mu_\theta^2} d\theta \tag{4.25}$$

It measures the path length weighted by the local critical sample sizes, scaled by the effective step size η_0. Figure 4.5 illustrates a non-convex setting, where the parameter is in a region of low gradient, leading to high learning times.

4.2.7 Convergence properties of Sampa

For typical convex problems, the objective function is quadratic around the optimum, so that the gradient increases linearly with the distance ϵ to the optimum, $|\mu| \propto \epsilon$. In general, the

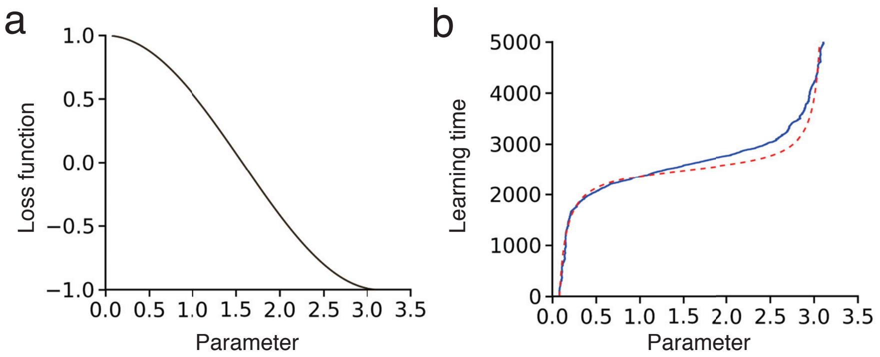

Figure 4.5 – **Learning time for SGD in non-convex setting.** (a) Non-convex loss function, $cos(\theta)$, with noise ($\sigma = 1.$) added to the gradient. (b) Learning time dependence on parameter evolution, for SGD simulation with Sampa, $\eta_0 = 0.01$ and $\theta_0 = 0.08$ (blue). Estimation theory prediction of learning time, c.f. Eq. 4.25 (red).

noise σ has no asymptotic scaling. Integration of Equation 4.25 gives a convergence scaling of

$$T \propto \frac{1}{\eta_0}\epsilon^{-1} \tag{4.26}$$

as long as $\epsilon > \eta_0$. Fig. 4.6b shows the convergence properties for three different values of η_0 (same convergence definition as in Fig. 4.4). Larger effective step sizes imply faster learning, but final convergence is scales with η_0. Fig. 4.6c illustrates the ϵ^{-1} scaling. However, for a given error ϵ target, estimation theory tells us we should set $\eta_0 \approx \epsilon$. Substituting this in Eq. 4.26, if we may choose the learning rate for a given target ϵ, we have a convergence time of $T \propto \epsilon^{-2}$.

4.3 Discussion

Our estimation theory for stochastic gradient descent is expected to have impact in both theory and practice of large-scale stochastic optimization. The analysis of gradient sampling statistics leads to a improved theoretical understanding of learning dynamics and convergence in SGD. In its practice, the theory and model of optimal adaptive learning rates should guide model choices in real-world applications.

4.3.1 Unifying theory for adaptive learning rates

Our results reveal a clear interpretation for each component commonly present in adaptive learning rates. Our theory emphasizes the importance of B^* as the minimum number of samples needed to estimate the gradient. In the following we compare the total parameter change $\Delta\theta^*$ over B^* samples for different adaptive models. Note that $\Delta\theta^*$ sets the scale of the standard deviation (jitter) of the trajectory of parameters (Fig. 4.3).

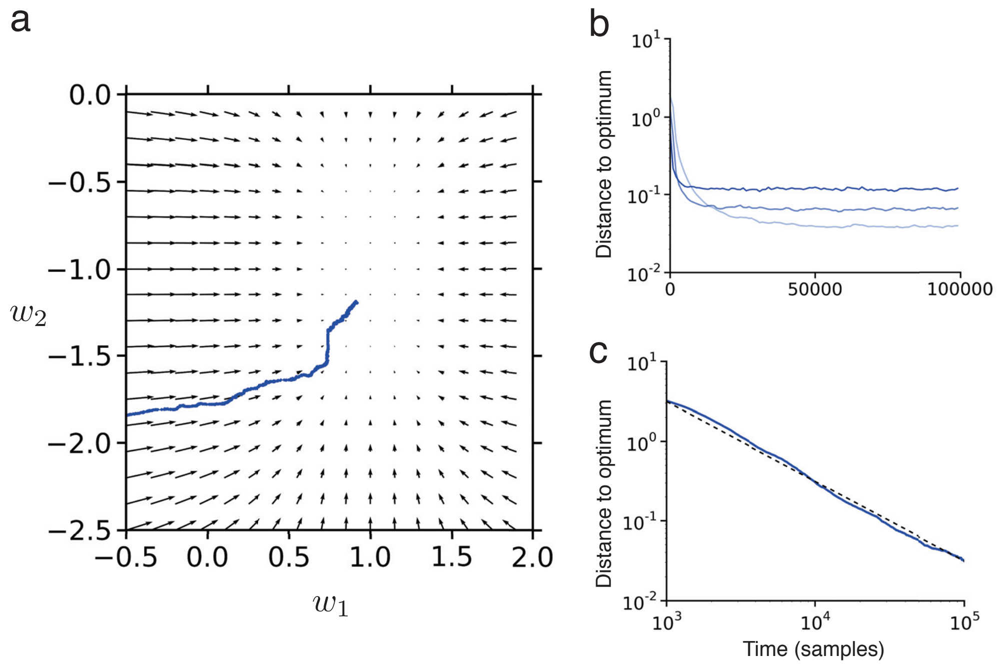

Figure 4.6 – **Learning dynamics and convergence properties for optimal SGD.** (**a**) The trajectory of the parameters for Sampa (blue) follow the velocity field given by the signal-to-noise ratio of the gradient ($\eta_0 = 0.01$). (**b**) Learning dynamics for Sampa for three different effective learning rates ($\eta_0 = 0.4$, 0.12 and 0.04; dark, medium and light blue respectively). Smaller effective step size implies slower learning but convergence to smaller distances. (**c**) Convergence of optimal SGD in quadratic optimization follows a T^{-1} scaling (blue, $\eta_0 = 0.01$). Dashed line is $c\ T^{-1}$ fit.

There are three main possibilities for adapting the learning rate to the statistics of the gradient samples.

- **No adaptation.**

  Standard SGD does not adapt to the statistics of the gradient samples. It implies that the the parameter change after B^* samples depends on the mean and variance of the gradient at each moment, $\Delta\theta^* \propto \frac{{\sigma^2}^t}{|\mu|^t}$ (cf. Eq. 4.9).

- **Variance normalization:** $\eta \propto \frac{1}{\sigma^*}$

  A simple form of sampling adaptation (e.g. Rmsprop) is to add a factor that normalizes the gradient's variability. While the magnitude of the parameter change taken after B^* samples becomes independent of the gradient's scale, we have shown that it is still modulated by the gradient's signal-to-noise ratio, $\Delta\theta^* \propto \frac{\sigma^t}{|\mu|^t}$ (cf. Eq. 4.9). The method requires the online estimation of σ^*.

- **Sampling normalization:** $\eta \propto \frac{|\mu|^*}{{\sigma^2}^*}$

> This factor cancels both the dependency upon the gradient's scale and signal-to-noise ratio, enabling optimal gradient estimation and a constant effective step size $\Delta\theta^* = c$ (cf. Eq. 4.11). This method (Sampa) requires the online estimation of σ^{2^*}, $|\mu|^*$, and an adapting time scale, $\tau_\mu = \frac{\sigma^2}{\mu^2}$.

In large-scale applications such as neural networks, which contain a large number of parameters and require long learning times, gradients may have diverse sampling statistics, across time and across parameters. Without adaptation to these statistics, parameter change may become too large or too small compared to the optimal step size. Variance normalization, as in Rmsprop, partially mitigates this problem by making the effective step size independent of the gradient scale, and since the online estimation of σ^* is relatively straight-forward, it is a robust method.

Our proposed method Sampa further increases independence from gradient statistics, which may lead to significant performance gains when the signal-to-noise ratio varies substantially across parameters or across time.

4.3.2 Analysis of previous adaptive models

Estimation theory enables a novel interpretation of the properties of previous adaptive models. Rmsprop is a widely used and simple adaptive model that implements variance normalization, $\eta = \eta_0 \, \frac{g}{\sigma^*}$ (Tieleman and Hinton, 2012). Closely related to Rmsprop is AdaDelta (Zeiler, 2012) (see Appendix).

Another model similar to Rmsprop is Adam, which implements $\eta = \eta_0 \, \frac{\bar{g}}{\sigma^*}$, thus maintaining a fixed effective step size (Kingma and Ba, 2014). But this model only implements variance normalization, being equivalent to Rmsprop with momentum, and it does not have the properties of Sampa. Adam has also been proposed to be more robust for sparse gradients than Rmsprop. However, this may be due to the choice of a smaller value than usual for τ_σ.

Adagrad is an adaptive method that incorporates a normalization factor with the norm of all previous gradients, $\eta^t = \eta_0 / \sqrt{\sum^t g_t^2}$ (Duchi et al., 2011). Under the assumption that the variance σ^2 does not change significantly in time, it can be rewritten as $\eta^t = \frac{\eta_0}{\sqrt{T}} \, \frac{1}{\sigma}$, thus approximating variance normalization with annealing ($\alpha = 0.5$).

Schaul et al. (2012) have analyzed the optimal update step size when parameters are close to a minimum. The adaptive model derived, $\Delta\theta = \frac{\mu^{2^*}}{\sigma^{2^*}} \, \frac{g_t}{h}$, where h is an estimation of the Hessian, can be reinterpreted by estimation theory. Near an optimum, the surface is quadratic, and Newton's methods tell us that μ/h is the optimal step size. Since the gradient is sampled through SGD, this step size must be rescaled by the critical sampling size, $1/B^*$, leading to the adaptive learning rate $\eta = \frac{1}{h} \, \frac{\mu^{2^*}}{\sigma^{2^*}}$. Thus their model implements signal-to-noise normalization with second-order optimization. Since it is optimal for convex optimization near the optima it may have good convergence properties in well-behaved convex problems, but a robust

	Statistics adaptation	Geometry adaptation	Rate decay	Averaging
Rmsprop	Variance	-	-	-
Adadelta	Variance	-	-	-
Adam	Variance	-	-	Momentum
Adagrad	Variance	-	Annealing	-
Schaul	SNR	Hessian	-	-
Sampa	SNR	-	-	Sampa (Eq. 4.21)

Table 4.1 – **Summary of various adaptive learning rate models in their standard implementations.** Properties such as momentum and annealing are in fact optional additions to all models.

estimate of the Hessian may not be possible in general and Newton's step may not be suitable for highly non-convex problems.

We conclude that estimation theory gives a clear framework for analyzing variations of adaptive learning rates in existing theories. In Table 4.1, we summarize the properties of each adaptive method in terms of four categories: adaptation to sampling statistics, adaptation to geometry, averaging, and rate decay.

Importantly, our analysis allows a practitioner to choose each property independently, setting up an adaptive learning rate tailored to the application at hand, not restricted to one of the previously proposed models.

4.3.3 Relation to previous plasticity models based on sampling

Recent studies have investigated how synapses could optimally adapt according to an objective function and the statistics of the neural activity (Aitchison and Latham, 2014; Kappel et al., 2015). In these studies the stochasticity in the synaptic strength is linked to the uncertainty about the optimal synaptic values. The more data is collected, more certain the synapse is of its optimal value, and the variability would decrease accordingly.

These models have intentions similar to the estimation theory in that they aim at optimizing the information extracted from stochastic gradients, as well as explaining how synapses may modulate their speed of change. However their assumption that variability is useful in representing uncertainty is opposite to our suggestion that variability is a property of stochastic optimization, to be overcome by adapting the effective sample size. Also these studies assume specific neural models and objective functions, so the resulting adaptive mechanisms and dynamics are particular to the network model under consideration.

Estimation theory and adaptive learning rate make no assumptions about the network model or objective function being optimized, and so it is applicable to any stochastic optimization problem. As such, our findings may appear as a component of solutions derived from more particular assumptions, as it is the case in Schaul et al. (2012). As the algorithms proposed in

Aitchison and Latham (2014) and Kappel et al. (2015) do not conform to the typical formulations of models analyzed here, it is not clear if they possess the adaptive properties we have analyzed.

4.3.4 Relation to geometry-based gradient descent studies

Studies of learning dynamics and convergence have been largely based on asymptotic convergence properties, for both offline and online algorithms (Bottou, 2010). For example, variants of Newton's method take into account the curvature of the optimization surface to estimate an appropriate update step size, and can be shown to have faster convergence properties near the optimum than standard gradient descent (Schaul et al., 2012). Another common approach relies on the implementation of annealing, in which the learning rate decreases with time, which also gives asymptotic guarantees on how close to the optimum the parameters will be in due time.

These approaches have important limitations. Newton's method depends on the estimation of the Hessian, which is computationally expensive and often erratic, although recent research has aimed at developing second-order models appropriate for non-convex settings (Dauphin et al., 2014; Gulcehre and Bengio, 2014).

Annealing theorems also have limited applicability since results are valid only asymptotically. These results only guarantee certain convergence rates, but make no statement on the large constants in the transient learning time that may come from the implementation of annealing.

We claim that in many applications it is the transient that is relevant: the period between the start and the moment when an approximate solution has been found with some acceptable precision. When facing large-scale complex optimization problems, asymptotic convergence may be the least of one's concerns.

4.3.5 Relation between gradient surface and estimation theory

We have shown that in SGD the learning dynamics are constrained by the gradient's signal-to-noise ratio surface. It is a novel perspective on the analysis of optimization problems, which so far have been focused on the geometry of the true gradient surface.

There are two facets to the relation between the gradient and the gradient's SNR. If the sampling noise σ does not change considerably, the SNR surface will have similar geometry to the true gradient surface (though determined by the square of the gradient). However, the crucial difference is that the gradient surface may be transformed for optimization purposes, for example by reparametrization, natural gradients or Newton's method. On the other hand, the SNR surface is a hard constraint for estimating the gradient, which cannot be trivially circumvented by such methods.

4.3.6 Non-convex high-dimensional optimization

Contemporary applications of machine learning, such as large neural networks, are non-convex problems. The presence of local minima and saddle-nodes in the gradient surface have been pointed as one of the main obstacles for the scaling up of these models (Saxe et al., 2013; Dauphin et al., 2014). When analyzing their learning dynamics, parameters seem to get stuck at some values for large periods of time, indicating very small gradients, and only with large amounts of data and time the parameters escape to further minimize the optimization function.

More than exceptions, these events are common, and it has been shown that there are exponentially many saddle-nodes in typical network models (Dauphin et al., 2014), raising interest in techniques that can escape them efficiently (Dauphin et al., 2014, 2015). Methods that have been proposed are based on the gradient geometry, as in Newton's method, using for example the absolute value of the curvature to determine efficient step sizes.

We suggest that possibly the slow learning dynamics due to small gradients may not be trivially circumvented by second-order methods. Near saddle-nodes the gradient is small, while the sampling noise is generally not significantly altered, implying low SNR. Thus many samples are needed to estimate the gradient and determine which direction to follow, allowing for a theoretical explanation for long learning times in large neural networks.

In the next chapter we show important implications of estimation theory using the SNR surface. By analyzing the gradient statistics and learning dynamics of a typical non-convex problem, we establish theoretical bounds on learning time and on the dimensionality of a network.

4.3.7 Implications for cortical plasticity

By interpreting cortical plasticity as a stochastic gradient descent process, the conclusions presented here give interesting predictions and suggestions about how synapses may regulate their weight change.

Since synaptic plasticity depends on the activity of the post- and pre-synaptic neurons, the gradient statistics will depend on the recent firing activity, which is known to vary across neurons. Artificial neural networks also have such properties, and have shown that optimizing the update step size can lead to significant performance gains in learning times.

As the brain has large incentives to optimize its learning efficiency, we believe that adaptive learning rates is an important mechanism to be investigated experimentally and that the adaptive models studied here are simple enough to be plausibly implemented in a synapse. And since no assumptions are made on the nature of the gradient signal or its correspondent objective function, our predictions are equally applicable to any plasticity mechanism, including variations of Hebbian learning, such as reward modulated plasticity.

Our theory is also insightful for models of discrete synaptic strengths and structural plasticity. In ongoing work, we are investigating an alternative version of Sampa with discrete updates, each update using adaptive batch sizes based on the critical sample size. This model allows for learning with continuous stochastic gradients, but sparse discrete changes, and may be simpler to implement by biological systems.

Finally, the application of estimation theory suggests surprising theoretical bounds on the duration of perceptual development and on the number of synapses that contact each neuron, as we demonstrate in the next chapter.

5 Limits of unsupervised feature learning in high dimensions

Unsupervised learning of a representation from complex high-dimensional inputs typically involves tuning a large number of parameters and requires vast amounts of data. A representation by multiple neurons with localized receptive fields has empirically been found useful in convolutional neural networks and computer vision, but a formal explanation for the advantages of local receptive fields is still lacking. Based on the geometry of high-dimensional spaces, we study how the input dimensionality of each neuron impacts the duration of feature learning. Our analysis exploits the fact that random directions are almost orthogonal to each other in high-dimensional spaces. This fact implies that random initial parameters have small overlap with hidden features, leading to small gradients and consequently to large learning times. Simulations confirm the theoretical predictions of a learning time with supralinear dependency on the input dimensionality. Our results explain why bounding the receptive field size is useful in practical applications. Our approach outlines a new framework for analyzing learning dynamics and model complexity in neural networks.

5.1 Introduction

Neural networks demand large amounts of data and learning time to develop representations (Krizhevsky et al., 2012). Recent breakthroughs in machine learning owe much of their success to the increasing availability of computing power and large-scale datasets (LeCun et al., 2015). Even with current resources, learning time remains one of the main obstacles in scaling up the complexity of neural networks and their applications (Le, 2013).

Both the number of neurons and the number of inputs per neuron (fan-in) influence learning time. Cortical networks, with their billions of neurons and thousands of synapses contacting each of them (Kandel et al., 2000), are likely to be limited by similar architectural constraints. The estimation theory developed in the previous chapter provides us with a powerful tool to analyze the learning dynamics in neural networks and get insights into how network complexity affects the time to convergence.

We consider the projection pursuit problem of finding hidden features in an input arriving through N synapses onto a neuron, where N determines the dimensionality of the optimization problem. We show that the optimization function of the projection pursuit problem has a number of saddle points and maxima that increases exponentially with the dimensionality. For a large number N of synapses, it becomes highly probable that random initial synaptic weights will be almost orthogonal to the hidden features. In other words, we claim that random initial conditions lead to an initialization close to saddle points, therefore falling in parameter regions of small gradients, which lead to large learning times.

Based on the geometry and statistical properties of high-dimensional spaces, we show that the learning dynamics are well described by a simple dynamical system, in which the only relevant variable is the initial distance to the closest hidden features which in turn increases with the number of synapses.

Our results lead to a striking analytical prediction: the optimal learning time has a supralinear dependency on the number of synapses onto a single neuron. Implicitly, this strong dependency induces an effective bound on synaptic connectivity, above which learning becomes unreasonably slow. We therefore speculate that the number of synapses per neuron in cortical networks is limited by learning constraints. Our results also provide a candidate explanation for the large performance gains observed in artificial neural networks such as convolutional networks that implement localized receptive fields and therefore a limit number of synaptic connections onto each neuron.

5.2 Results

5.2.1 Unsupervised projection pursuit learning in neural networks

We consider a network of M independent neurons, where each neuron receives N synapses. We assume that synaptic plasticity implements the minimization of an objective function. For each neuron in the network, synapses change according to the gradient of this function. To be specific, we consider the projection pursuit problem of finding hidden features in the neuron's input.

In projection pursuit, we search for hidden patterns by optimizing synaptic weights $\mathbf{w}$ according to some optimization function,

$$\min_{\mathbf{w}} \left\langle -F(\mathbf{w}^T \mathbf{x}) \right\rangle \tag{5.1}$$

where we define it as a minimization problem for consistency with the previous chapter. We model the N-dimensional input $\mathbf{x}$ to the neuron as K independent hidden features $x_i \sim \text{Laplace}(0, 1/\sqrt{2})$, complemented by $N-K$ Gaussian inputs, defining the input probability distribution

$$P(\mathbf{x}) \propto \prod_{i=1}^{K} e^{-\sqrt{2}\,|x_i|} \prod_{i=K+1}^{N} e^{-\frac{1}{2}x_i^2} \tag{5.2}$$

We constrain the norm of the weights to $|\mathbf{w}|_2 = 1$, thus optimizing only the direction of $\mathbf{w}$. We also note that the optimization function has no dependency upon second-order statistics, since the input defined in Eq. 5.2 is white, $\langle \mathbf{x}\mathbf{x}^T \rangle = I$. The aim of projection pursuit is to align the direction of the weight vector $\mathbf{w}$ with one of the N hidden features. As discussed in Chapter 2, many possible optimization functions have optima at the Laplacian hidden patterns (Hyvarinen and Oja, 1998). Implementing projection pursuit by stochastic gradient descent leads to the learning rule, $\mathbf{g}_t = -\mathbf{x}_t\, f(\mathbf{w}^T\mathbf{x}_t)$, where $F(z) = \int_0^z f(x)dx$. Throughout our analysis we use a linear rectifier with slope 1 and threshold 3 as nonlinearity, $f(u) = (u-3.)_+$.

5.2.2 The geometry of the optimization surface in neural networks

To understand the learning behavior of our model, we start by developing an analytical description of the properties of our optimization surface. We determine the position and quantity of minima, maxima, and saddle points, which are the critical points of gradient descent methods because the gradient is zero. Saddle-points have zero derivative, but present positive curvature in some parameter directions, while negative in others. Figure 5.1a-c illustrates the the typical geometry of these special points on an optimization surface in two dimensions.

Given the input distribution defined in Eq. 5.2 with $K = N$, the optimization function has a minimum w^* at each of the N hidden features, which are the cardinal directions in our example, $w^* = (0,\ldots,0,\pm1,0,\ldots,0)$, yielding a total of $2N$ minima. At the same time, there are exponentially many maxima, at each of the 2^N symmetric directions, $w^{max} = (\pm1,\pm1,\ldots,\pm1,\pm1)/\sqrt{N}$.

In addition, even more saddle points than maxima exist. Although hard to visualize, any direction with k symmetric components $(1 < k < N)$, $|w_{i_1}| = \ldots = |w_{i_k}| \neq 0$, while other components are zero, is a saddle point. It is a maximum in respect to the k non-zero components, and a minimum in respect to the other $N-k$ null components. We show in the Methods that it implies a total number on the order of 3^N saddle points.

Figure 5.1d illustrates these properties for the three-dimensional case. The 6 cardinal directions represent the minima, where the hidden features lie. Each one of these is surrounded by a basin of attraction. The 8 symmetric directions are maxima, while the 12 partially symmetric points are the saddle points.

The parameter regions that connect maxima and saddle points have small gradients throughout and are far from the minima. Starting at these regions leads to large learning times. Note that the maxima and saddles exist due to the symmetries in the problem. Saddle points due to symmetries have previously been observed in other network models (Saxe et al., 2013; Choromanska et al., 2014; Dauphin et al., 2014). The numerical predominance of saddle points is also in alignment with these other studies (Saxe et al., 2013; Choromanska et al., 2014; Dauphin et al., 2014). In contrast to these earlier results, our approach uses mainly geometric arguments in the framework of projection pursuit defined in Eqs. 5.1 and 5.2.

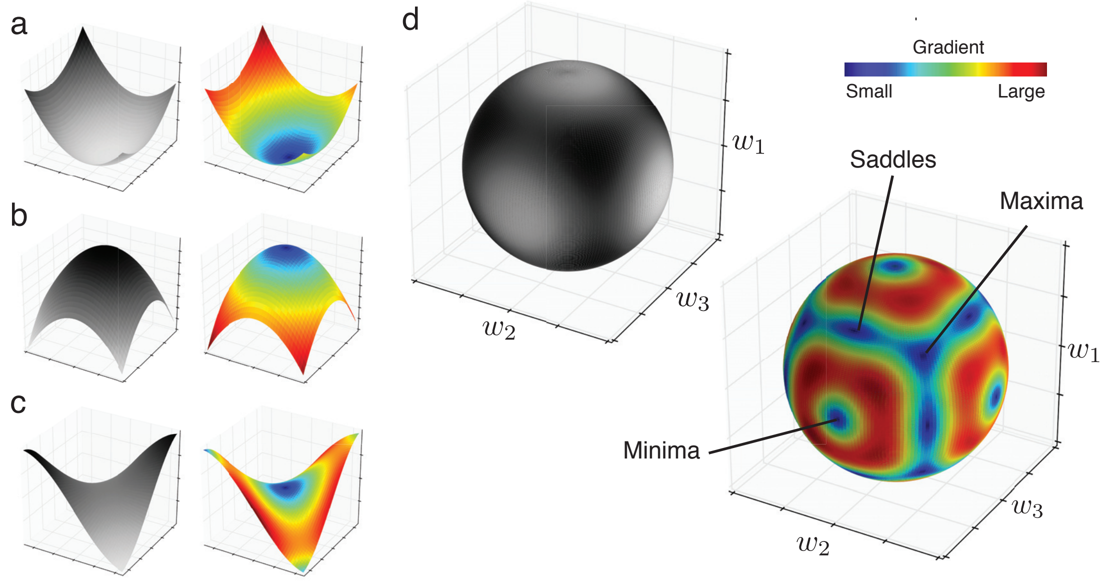

Figure 5.1 – **Geometry of the optimization surface for synaptic weights.** (**a**) Prototypical minimum in a convex surface with a basin of attraction. (left) Gray scale indicates optimization function value, with minimum in white. (right) Color heat map indicate the magnitude of the gradient, with zero gradient in dark blue. (**b**) Prototypical maximum in a concave surface, with an unstable equilibrium point. (**c**) Saddle points are concave in some parameter directions, while convex in others. (**d**) Surface for the gradient magnitude for three synapses. Each dimension represents the value of a synaptic weight. As we enforce $|\mathbf{w}|_2 = 1$, the parameter space is constrained to the unit sphere. The cardinal directions are the minima, the directions of the hidden patterns. The symmetric directions, where all weights have the same magnitude, are maxima. Partially symmetric directions, where two weights have the same magnitude, and the third is zero, are saddle points. Areas in blue indicate low gradient magnitude, and thus slow learning dynamics.

5.2.3 Quasi-orthogonal random directions in high-dimensional spaces

The three dimensional example illustrated above gives a good intuition for some general properties of the optimization surface of feature learning by projection pursuit. However, high-dimensional spaces have anti-intuitive geometrical properties that are not evident in low dimensional cases. The essential property that we will exploit for our analysis is the angular distance between random vectors.

We note that the initial values of the synaptic weights are, in most practical applications, randomly distributed. Before we turn to random vectors, let us consider a "worst case" scenario.

In the three dimensional case, any vector will be at most at 55° angular distance to one of the cardinal directions, with the maximal distance at the maxima (worst case). However, in higher dimensions this distance increases asymptotically to 90°, with the overlap between minima and maxima in N dimensions being $d_{\max} = \mathbf{w}_{\max}^T \mathbf{w}^* = 1/\sqrt{N}$ (see Methods).

Importantly for our purposes, random directions $\mathbf{w}^R$ follow a similar decay in their average distance d to the closest minimum, $d \approx \frac{\sqrt{2\log(N)}}{\sqrt{N}}$ (see Methods). This result implies that initial parameters will have only a small overlap with the hidden features, starting from almost orthogonal angles.

Although we have so far assumed the same number of hidden features K and input dimensions, $K = N$, our results can also be adapted to $K < N$ and $K > N$ (see Methods), leading to an overlap given by

$$d^K \approx \frac{\sqrt{2\log(K)}}{\sqrt{N}} \tag{5.3}$$

Thus, for large N, the expected overlap will still be small even if when the number of hidden features is large in relation to the number of dimensions ($K > N$). It is a by-product of the rather exceptional fact that in high dimensions one may select exponentially many random vectors, and all of them will be almost orthogonal (namely quasi-orthogonal) to each other with high probability (Cai et al., 2013).

In Figure 5.2 we illustrate these dependencies in a simulation with randomly generated directions. Figure 5.2b shows how, given an input dimensionality N, the initial overlap d^K increases only logarithmically with the number of hidden features.

We can relate these findings to the intuition gained in the three dimensional case. When the number of dimensions increases, the area of the blue region filled with saddles and maxima (see Figure 5.1d) becomes exponentially larger than the area composed by the basins of attraction around the minima. Formally, with higher dimensions the area of the polar cap around a given direction decreases exponentially in comparison to the total area of the sphere (Li, 2011).

These results imply that when there are many input synapses, the initial random weights will be quasi-orthogonal to hidden features, and lie in the parameter region of small gradients that is filled with saddle points and maxima.

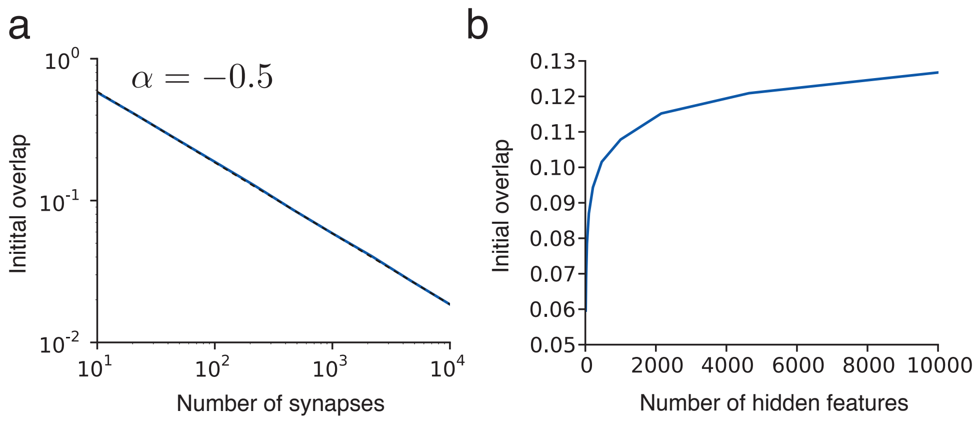

Figure 5.2 – **Initial parameters have small overlap with hidden feaures in high dimensions.** (**a**) The expected overlap of initial parameters with a hidden feature decays with the square root of the number of dimensions N (fixed $K = 10$). Dashed line is power-law with exponent $\alpha = -0.5$. (**b**) Increasing the number of hidden features has only a logarithmic effect on the expected initial overlap (fixed $N = 1000$).

5.2.4 Statistical properties of high-dimensional inputs

Once we have established the general properties of the gradient values in terms of critical points, we can now analyze how the gradient changes in between the critical points. Following the results of the previous section, we assume a large dimensionality N and that the initial parameters have small overlap with the hidden features, with $K = N$.

Returning to our projection pursuit problem, the total input to the neuron is a mixture of the N hidden patterns, weighted by the synaptic strengths, $u = \mathbf{w}^T\mathbf{x} = w_1 x_1 + \cdots + w_N x_N$. Since for large N the initial individual weights w_i^0 are small, we may invoke the central limit theorem, and conclude that, as N increases, the total input $u(\mathbf{x})$ converges to a normal distribution.

We make the assumption that during learning the weights will converge to the closest hidden pattern $\mathbf{w}^* = \mathbf{e}_j = (0,\ldots,0,1,0,\ldots,0)$, following the shortest path from the initial weights $\mathbf{w}^0 = (w_1^0,\ldots,w_{j-1}^0, w_j^0 = d_0, w_{j+1}^0,\ldots,w_N^0)$, where j is the feature with maximal overlap and d_0 is the initial overlap.

In the following we do not consider the setting of stochastic online updates, but look at the full statistical distribution over infinitely many patterns. We rewrite the input as $u = w_j x_j + \sum_{j\neq k} w_k\, x_k$, and invoke again the central limit theorem, arriving at the approximation $u = w_j x_j + \sum_{j\neq k} w_k\, g = w_j\ l + \sqrt{1 - w_j^2}\ g$, where l is a Laplacian variable and g is a normally distributed variable, and we used the fact that the weights are normalized, $\sum w_j^2 = 1$.

Thus along the path from $\mathbf{w}^0$ to $\mathbf{w}^*$ we can approximate the change on the statistical distribution of u as a linear transition between the initial (normal) distribution and the final

(Laplacian) distribution,

$$u_t = d_t\, l + \sqrt{1 - {d_t}^2}\; g \tag{5.4}$$

where we parametrize the parameter evolution by the overlap d.

The formulation in Eq. 5.4 is revealing, since it shows that the distribution of inputs u_t only depends on the overlap d_t. It implies that the gradient of $\langle F(u) \rangle$ depends only on d_t, allowing us to study the learning dynamics through a simple one dimensional system, $\hat{F}(d_t) = \langle F(u_t) \rangle$, where u_t is given by Eq. 5.4. Gradient descent on this one-dimensional function gives (see Methods)

$$\Delta d_t \propto -\frac{\partial \hat{F}(d_t)}{\partial d} \tag{5.5}$$

In order to check whether the full N-dimensional stochastic projection pursuit can be represented by the effective one-dimensional dynamics (Eq. 5.5), we investigate the learning dynamics by simulating our projection pursuit problem through stochastic gradient descent for a variety of dimensions, implementing gradient descent, $\Delta \mathbf{w}_t = \eta\, g_t$, and starting from random weights $\mathbf{w}_0$. Figure 5.3a shows the time evolution of the largest overlap d_t between $\mathbf{w}_t$ and a hidden feature. It shows that initial overlaps depend on the dimensionality, with higher dimensions tipically implying lower initial overlap, and thus larger learning time.

It also indicates that all runs have similar trajectory profiles, which is expected from our reduction of the learning dynamics to a single one dimensional system, that only depends on N through the initial overlap d_0. In Figure 5.3b we highlight the stereotypical learning dynamics by showing the trajectories with a shifted time reference, to the point when the overlap crossed $d_t = 0.75$.

5.2.5 Learning time is constrained by the initial overlap

By analyzing the gradient properties, we may obtain analytical expressions for the learning dynamics that we have observed. As we have shown in the previous chapter, the optimal learning dynamics is dictated by the signal-to-noise ratio of the gradient. In Figure 5.4 we show the gradient statistics dependent on the overlap d, by simulating our reduced unidimensional model of Equation 5.5. We see that the gradient SNR goes quickly to zero for smaller overlaps (Figure 5.4c). Also, the SNR profile follows the profile of the gradient magnitude (Figure 5.4a), since the gradient variability does not significantly change for small overlaps (Figure 5.4b).

In the previous chapter we showed how to estimate the total learning time based on the gradient statistics, from a starting point d_0 to a target d^*,

$$T(d_0) = \frac{1}{\eta_0} \int_{d_0}^{d^*} \frac{\sigma^2_{\tilde{d}}}{\mu^2_{\tilde{d}}} \partial \tilde{d} \tag{5.6}$$

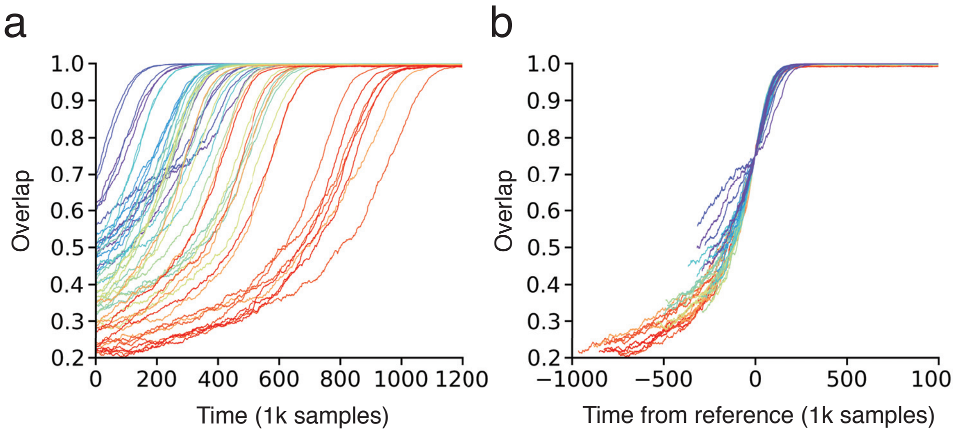

Figure 5.3 – **Stereotypical learning dynamics in high-dimensional optimization.** (**a**) Evolution of the overlap of the weights with a hidden feature for different number of synapses. The color heat indicate N, varying from $N = 10$ (purple) to $N = 160$ (red). One run for each N, with random initial weights. (**b**) Same trajectories shifted to a referent time where an overlap of $d = 0.75$ was reached, highlighting the similarity in the learning dynamics.

where we assume a constant effective step size η_0, $\sigma^2_{\tilde{d}}$ and $\mu^2_{\tilde{d}}$ evaluated using Eq. 5.4. This expression allows us to predict the learning dynamics seen in the parameter trajectories in Figure 5.3b.

For simplicity, we consider a constant gradient variability $\sigma = c$, leaving us to find out how the gradient magnitude $|\mu|$ depends on the overlap d. In Methods, we derive an analytical expression for how the gradient magnitude $|\mu|$ depends on the overlap d asymptotically, by analyzing the statistical properties of the gradient for small overlaps. Intuitively, the derivation follows from an analysis of statistical moments. The first three moments of F are constant, since the odd moments are zero for symmetric distributions and the second order is the variance, which is constant for white input. Thus the fourth-moment is smallest one to depend on d, which leads to a forth order dependency for F, and consequently a third order dependency for the gradient,

$$\hat{F}(\mathbf{d}) \propto a + d^4 \implies \mu(\mathbf{d}) \propto d^3 \tag{5.7}$$

In Figure 5.4d, we confirm this scaling law by calculating the gradient for small overlaps.

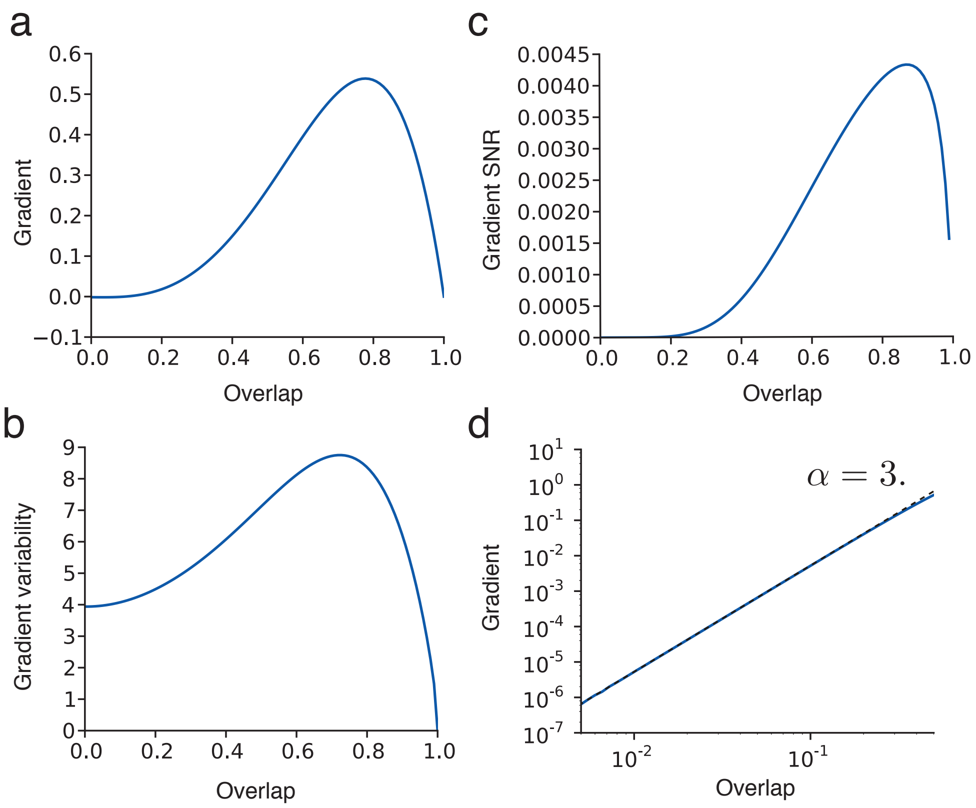

Figure 5.4 – **Gradient dependency on the overlap *d*.** (**a**) The gradient magnitude vanishes when the overlap goes to zero. (**b**) The gradient variability does not significantly change for small overlaps. (**c**) The gradient signal-to-noise ratio follows the profile of the gradient magnitude. (**d**) Near zero, the gradient has a power-law dependency on the overlap (dashed line is power-law with exponent $\alpha = 3.$).

5.2.6 Learning time dependency on number of synapses

We have succeeded to derive a chain of dependencies, in Eqs. 5.3, 5.6 and 5.7, that together reveal that the learning time has a supra-linear dependency on the input dimensionality,

$$\left.\begin{aligned} &d_0 \propto N^{-0.5} \\ &\mu \propto d^3 \\ &T \propto \int_{d_0}^{d^*} \mu_{\tilde{d}}^{-2}\, \partial\tilde{d} \end{aligned}\right\} \implies T \propto N^{2.5} \tag{5.8}$$

In Figure 5.5, we show that our gradient descent simulations confirm a learning time that scales with the number of synapses with a 2.5 exponent, and a proportionality factor around 5: $T \approx 5\, N^{2.5}$.

These results offer for the first time an analytical description of how the number of input synapses affects learning. It suggest a fundamental bound on the number of synapses during development. For instance, if we consider a budget $T = 10^9$ data samples, in our model the number of synapses must be lower than 2000.

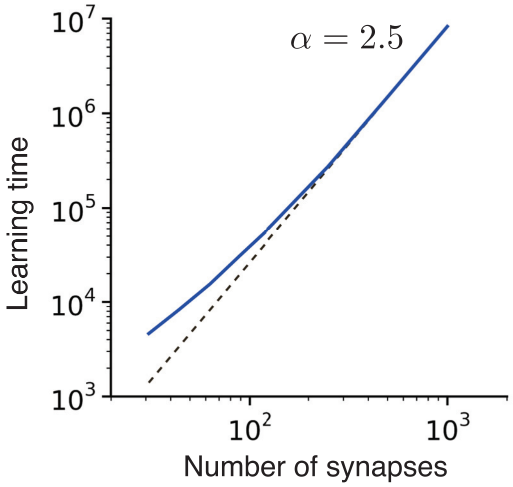

Figure 5.5 – **Learning time dependence on number of synapses.** Learning time depends on input dimensions N through a power-law with exponent 2.5 (dashed line is power-law with $\alpha = 2.5$). Learning time was defined as the average time to reach an overlap of $d = 0.75$. The asymptotic theory is valid for $N > 200$.

5.3 Discussion

5.3.1 Generalization of results to other neural network models

We have studied a particular model of unsupervised learning. However, the assumptions required for our results may be more general, and possibly transfered to other learning scenarios. For any network model in which the total input to the neuron is a linear projection, $u = \mathbf{w}^T\mathbf{x}$, the gradient for the synaptic strengths can be formalized as $g = \partial_w \left\langle F(\mathbf{w}^T\mathbf{x}) \right\rangle_x$, as in our study. Reinforcement learning or supervised learning networks are important examples.

The crucial assumption made for our unsupervised learning scenario is that the optimization function only depends on higher-order statistics of the input. Future investigations shall determine if we can extend the approach to alternative learning paradigms, where target values or rewards also play a role.

5.3.2 Theory for convolutional neural networks

Large hierarchical neural network models have recently obtained impressive results in a variety of artificial inteligence tasks (LeCun et al., 2015; Bengio et al., 2013). A large part of these applications rely on the implementation of convolutional neural networks, in which each neuron has a limited receptive field size (Krizhevsky et al., 2012; Mnih et al., 2013). This architectural constraint is essential when the input has high dimensionality, as in the case of images.

Despite its popularity and proven efficiency, a definite theoretical explanation for the functional gains of limited input dimensions have been elusive. Current explanations rely on the heuristic arguments of information content, stating that less model parameters require less data. These arguments may be rebutted by the fact that larger data dimensionality also mean higher information content available. In any case, these arguments do not lead to an analytical understanding of how connectivity affects learning time.

Our results suggest a potential explanation for the performance gains due to localized receptive fields and connectivity constraints, showing analytically that larger receptive field sizes can make learning impractical. In ongoing work, we are investigating if the proposed scaling laws apply to large-scale convolutional networks.

5.3.3 Implications for neural network learning dynamics

The unsupervised learning paradigm we have used has proven to be a useful prototype in understanding how learning unfolds in neural networks. It allows us to calculate precisely the number of extreme points, such as saddle points, how they are distributed and how they determine the learning dynamics.

Our curvature analysis is aligned with previous studies that have collected evidence that saddle points are abundant and may be an important obstacle for neural network learning (Saxe et al., 2013; Dauphin et al., 2014; Choromanska et al., 2014). In Dauphin et al. (2014), the authors use numerical simulations to probe the existence of saddle points in large networks. Our results may help providing an analytical explanations for their findings.

In Saxe et al. (2013), the authors study a multi-layer linear network model, where they use the optimization gradient to qualitatively characterize the learning dynamics. However, the choice of a linear model led to different conclusions, and their results depend on properties of multi-layer architectures, but not on the learning properties of each neuron or the number of synapses. In their model, a single layer network does not have saddle points, and the symmetries in the optimization were due to the multiple layers, providing a complementary source of symmetries in large neural networks. Our analysis allows the study of nonlinear multi-layer networks, and in ongoing work we investigate how our results generalize to multiple layers.

The slow learning dynamics due to the presence of saddle-nodes in the gradient curvature have aroused interest in trying to overcome these obstacles through second order methods (Dauphin et al., 2014, 2015). However our estimation theory presented in Chapter 4 shows that it is the gradient signal-to-noise ratio that constrains learning. These limitations cannot be trivially overcome by curvature approaches.

5.3.4 Theory for localized receptive fields and number of synapses in the cortex

Cortical neurons have in the order of 1000 to 10000 input synapses and as space is a strong constraint in the brain, it is believed connectivity properties may be determined by the trade-off between representational power and volume (Rivera-Alba et al., 2011).

Our results provide an unexpected alternative explanation for synaptic densities. Learning is effectively bounded by the number of input synapses per neuron, and even if space was not a constraint, there would still be a limit on the order of a few thousand of synapses imposed by synaptic development. Note that different synapse types can have different roles and a subset of synapses might be non-plastic during development.

Although it is difficult to propose a direct test for this hypothesis, indirect evidence may support it. High-performance artificial neural networks constrain their connectivity by limiting receptive field sizes, without which their impressive performances are not possible (Le, 2013; Lee et al., 2009). It is a clear example where connectivity bounds are a functional requirement for developmental learning.

5.3.5 Implications in the duration of sensory development

Our theory also makes predictions about the time scales and learning dynamics in sensory development. As we have seen in Chapter 2, the development of receptive fields can be modeled as an optimization of synaptic connections. In this case, given the statistics of the neural activities, and models of synaptic plasticity, we may use our theory to estimate how much sensory data is needed for learning. It would provide a theoretically grounded explanation for the time scales of critical periods in sensory development (Berardi et al., 2000).

5.4 Methods

5.4.1 Number of maxima and saddle points

We provide here a short explanation for how to count the maxima and minima in the synaptic optimization surface. Maxima lie at the edges of a hypercube, where all weights have the same magnitude, $w_i = \pm 1/\sqrt{N}$. For each weight there are two possibilities, negative or positive, amounting to 2^N combinations.

Saddles have some zero weights, while the others have the same magnitude, $|w_{i_1}| = \cdots = |w_{i_k}| \neq 0$. Thus each weight has three possibilities, positive, negative or zero, amounting to 3^N combinations. The exceptions are the points that are minima, maxima, and the zero vector $\mathbf{w} = 0$, totaling $3^N - 2^N - 2N - 1$ saddle points.

5.4.2 Data generation

In general, the N-dimentional input $\mathbf{x}$ to our unsupervised learning problem is generated by a whitened linear mixture of K Laplacian variables, $\tilde{\mathbf{x}} = \sum_{i=1}^{K} \mathbf{w}_i l_i$, where $l_i \sim \text{Laplace}(0, 1/\sqrt{2})$ and the mixing vectors $\mathbf{w}_i$ are N-dimensional.

For $K = N$, we used orthogonal mixing vectors, $\mathbf{w}_i = (0, \ldots, 0, w_i = 1, 0, \ldots, 0)$, which generates white inputs and defines the probability distribution

$$P(\mathbf{x}) \propto \prod_{i=1}^{N} e^{-\sqrt{2}\,|x_i|} \tag{5.9}$$

For $K < N$, we add Gaussian variables to complete N dimensions, generating the input by $\mathbf{x} = \sum_{i=1}^{K} \mathbf{w}_i l_i + \sum_{i=K+1}^{N} \mathbf{w}_i g_i$, where $g_i \sim \text{Normal}(0,1)$, and same mixing vectors as in the previous case. It defines the probability distribution

$$P(\mathbf{x}) \propto \prod_{i=1}^{K} e^{-\sqrt{2}\,|x_i|} \prod_{i=K+1}^{N} e^{-\frac{1}{2}x_i^2} \tag{5.10}$$

For $K > N$, we generate the input by $\tilde{\mathbf{x}} = \sum_{i=1}^{K} \mathbf{w}_i l_i$, for random vectors $\mathbf{w}_i$ with $|\mathbf{w}_i| = 1$ (see below). The resulting data is then whitened, $\mathbf{x} = \mathbf{M}\,\tilde{\mathbf{x}}$ (see 2.8).

5.4.3 Expected overlap between random vectors in high dimensional spaces

The average overlap between two random directions can be derived from how random directions $\mathbf{w}^R$ are generated in an N-dimensional sphere: each component is drawn from a normally distributed variable, and the resulting vector is normalized to unit norm,

$$\tilde{\mathbf{w}} = (\tilde{w}_1, \ldots, \tilde{w}_N),\ w_i \sim N(0,1) \tag{5.11}$$

$$\mathbf{w}^R = \frac{\tilde{\mathbf{w}}}{|\tilde{\mathbf{w}}|} \tag{5.12}$$

Since for large N the norm $|\tilde{\mathbf{w}}|$ is well approximated by $\sqrt{N}$, we conclude that each component of the random vector follows $w_i^R \sim N(0, \frac{1}{\sqrt{N}})$. If we consider without loss of generality that the reference direction is $\mathbf{w}^* = (1, 0, \ldots, 0)$, the overlap will follow a distribution $d \sim N(0, \frac{1}{\sqrt{N}})$.

Considering K random directions $\mathbf{w}^k$ instead, the expected largest overlap d^K to a reference direction, $\mathbf{w}^* = (1, 0, \ldots, 0)$, is given by the maximal value amongst the K first components, $d^K = max_k\ w_1^k \approx \frac{\sqrt{2\log(K)}}{\sqrt{N}}$, where we used the expected extreme value amongst K normally distributed variables for large K, $max_k\ g_k \approx \sqrt{2logK}$ (David and Nagaraja, 1970).

It follows that if there are K random hidden features in an N-dimensional space, the initial random parameters will have an expected largest overlap of approximately $\frac{\sqrt{2\log(K)}}{\sqrt{N}}$ to one of the hidden features.

5.4.4 Gradient scaling law for small overlaps

We wish to analyze the gradient magnitude at small overlaps, $d \to 0$. We start by analyzing the statistical properties of the projected input u, through its cumulants, where $u = d\,l + \sqrt{1-d^2}\,g$. The odd cumulants κ^u_{2m+1} are zero since u has a symmetric distribution. The second cumulant is constant, $\kappa^u_2 = 1$, since the variance is constant. Higher cumulants depend only on l, $\kappa^u_{2m} = \kappa^l_{2m} d^{2m}$, for $m > 1$, since the normal variable g has null higher order cumulants, $\kappa^g_{2m} = 0, m > 1$.

We calculate the Taylor expansion of the optimization function at $u = 0$,

$$\langle F(u)\rangle = \langle F(0) + uF^{(1)}(0) + u^2F^{(2)}(0)/2 + u^3F^{(3)}(0)/3! + u^4F^{(4)}(0)/4! + \dots\rangle \tag{5.13}$$

$$= a + \frac{F^{(4)}(0)}{4!}\langle u^4\rangle + \dots = a + \frac{F^{(4)}(0)}{4!}\,\kappa^l_4\,d^4 + O(d^6) \tag{5.14}$$

$$\approx a + b\,d^4 \tag{5.15}$$

$$= \hat{F}(d) \tag{5.16}$$

for some constants a and b, showing that the optimization function has a forth order dependency on d.

Importantly, Eq. 5.15 enables us to formally reduce the gradient ascent to one dimension. Let $\mathbf{w}_j = e_j$ be the closest hidden feature to the initial weights, such that the overlap is given by the j-th component of the weights, $d = |w_j|$, and we assume $w_j > 0$ without loss of generality. We have that

$$\Delta w_j \propto \frac{\partial\left\langle F(\mathbf{w}^T\mathbf{x})\right\rangle}{\partial w_j} \implies \Delta d \propto \frac{\partial \hat{F}(d)}{\partial d} \tag{5.17}$$

so that the gradient for the overlap has a third order dependency at small overlaps,

$$\hat{F}(d) \propto a + d^4 \implies \Delta d \propto d^3 \tag{5.18}$$

6 Conclusion

We have presented novel theoretical and modeling results, which aim at an improved understanding of how neural circuits self-organize in order to represent sensory information. We have assumed throughout that neural networks are designed for learning the statistical regularities in its input, and each chapter has addressed an obstacle faced by such processes.

The common methodology across studies was to search for the fundamental principle behind each problem, and thus present hopefully general and simple solutions. Many previous models have investigated issues presented here, such as the functions of Hebbian plasticity models (Intrator and Cooper, 1992; Hyvarinen and Oja, 1998; Pfister et al., 2006; Zylberberg et al., 2011) or adaptive learning rates for stochastic optimization (Duchi et al., 2011; Schaul et al., 2012; Kingma and Ba, 2014), and our results have proposed an unification of these studies under a single perspective, by reinterpreting them as examples of a general class of models.

The final results can be summarized concisely through their most important and conclusions and formulas:

1. A large class of plasticity models, namely nonlinear Hebbian learning models, are able to learning receptive fields from natural inputs:

   $$\Delta \mathbf{w} \propto \mathbf{x}\, f(\mathbf{w}^T \mathbf{x})$$

2. Synaptic depression with meta-plasticity enables learning to be invariant to second-order statistics:

   $$\Delta \mathbf{w} \propto \mathbf{x}\, y^2 - h\, \mathbf{x}\, y$$

3. Learning dynamics in stochastic gradient descent is constrained by the signal-to-noise

ratio of the gradient samples:

$$B^* = \frac{\sigma^2}{\mu^2}$$

4. Learning time effectively bounds the number of synapses per neuron:

$$T \propto N^{2.5}$$

The simplicity of these results allows them to be easily interpretable and applicable to different models. Importantly, simplicity increases the possibility of their implementation by biological systems.

The studies presented here can be regarded as neurocentric, investigating the functional role of learning in individual neurons. Further work shall aim at reconciling these results with how populations of neurons represent information. Particularly, the role of recurrent connections, top-down modulation or lateral inhibition should be important in advancing the functional understanding of synaptic development and sensory representations.

A Adaptive model Adadelta is an approximation of Rmsprop

Adadelta is an adaptive learning rate model for stochastic gradient descent designed to implicitly approximate the Hessian of the gradient, thus performing second-order optimization (Zeiler, 2012). Here we demonstrate that the proposed model does not have second-order optimization properties, and is instead approximating Rmsprop (Tieleman and Hinton, 2012), implementing variance normalization.

The Adadelta learning rule is implemented as

$$\Delta x_t = \frac{\sqrt{\overline{\Delta x^2}_{t-1} + \epsilon}}{\sqrt{\overline{g^2}_t + \epsilon}}\, g_t \tag{A.1}$$

where g_t is the gradient at time t, x_t is the paramter update, $\epsilon \ll 1$ is a small constant, $\overline{\Delta x^2}$ and $\overline{g^2}$ are moving averages of the second-order moment of the update and gradient respectively, with a time scale $1/\rho$,

$$\overline{g^2}_t = (1-\rho)\,\overline{g^2}_{t-1} + \rho\, g_t^2 \tag{A.2}$$

$$\overline{\Delta x^2}_t = (1-\rho)\,\overline{\Delta x^2}_{t-1} + \rho\, \Delta x_t^2 \tag{A.3}$$

A.1 Analytical derivation

As in Chapter 4, we assume small updates and high gradient variability, allowing the central limit theorem approximation for the gradient, $g_t \sim N(\mu, \sigma)$.

We define the normalized gradient, $\xi_t = \frac{g_t}{\sqrt{\overline{g^2}_t + \epsilon}}$, and derive how the update variability depends on it,

$$\begin{aligned}
\overline{\Delta x^2}_t &= (1-\rho)\ \overline{\Delta x^2}_{t-1} + \rho\ \Delta x_t^2 \\
&= (1-\rho)\ \overline{\Delta x^2}_{t-1} + \rho\ (\overline{\Delta x^2}_{t-1} + \epsilon)\ \xi_t^2 \\
&= (1-\rho+\rho\ \xi_t^2)\ \overline{\Delta x^2}_{t-1} + \rho\ \epsilon\ \xi_t^2
\end{aligned}$$

$$\Longrightarrow \left\langle \overline{\Delta x^2}_t \right\rangle = \frac{\rho\ \epsilon\ \langle \xi_t^2 \rangle}{\rho - \rho\ \langle \xi_t^2 \rangle} = \frac{\epsilon}{\frac{1}{\langle \xi_t^2 \rangle} - 1} \tag{A.4}$$

We proceed to approximate the distribution of $\langle \xi_t^2 \rangle$,

$$\begin{aligned}
\langle \xi^2 \rangle &= \frac{g_t^2}{\overline{g^2}_t + \epsilon} \\
&= \frac{g_t^2}{(1-\rho)\overline{g^2}_{t-1} + \rho g_t^2 + \epsilon} \\
&\approx \frac{\sigma^2 \chi}{(1-\rho)\sigma^2 + \rho\sigma^2\chi + \epsilon}
\end{aligned} \tag{A.5}$$

where we approximate $\overline{g^2}_t = \sigma^2$ as a constant and $g_t^2 = \chi\sigma^2$ as a scaled chi-squared variable, where χ is a chi-square(1) variable. Assuming $\rho \ll 1$, we may rewrite,

$$\begin{aligned}
\langle \xi^2 \rangle &= \frac{1}{\rho + \frac{((1-\rho)+\epsilon/\sigma^2)}{\chi}} \\
&\approx \frac{1}{\rho + \frac{(1+\epsilon/\sigma^2)}{\chi}} \\
&= \frac{1}{\rho}\ \frac{1}{1 + \frac{(1+\epsilon/\sigma^2)/\rho}{\chi}}
\end{aligned} \tag{A.6}$$

The Taylor expansion of $f(c) = \langle \frac{1}{1+1/c\chi} \rangle$ gives an approximation $f(c) \approx c - 2c^2$ for small c, leading to

$$\begin{aligned}
\langle \xi^2 \rangle &\approx \frac{1}{\rho}\ (\frac{1}{1+\epsilon/\sigma^2}\rho - 2\frac{1}{(1+\epsilon/\sigma^2)^2}\rho^2) \\
&= \frac{1}{1+\epsilon/\sigma^2} - 2\ \frac{1}{(1+\epsilon/\sigma^2)^2}\ \rho \\
&= \frac{1}{(1+\epsilon/\sigma^2)^2}\ (1 - 2\rho + \epsilon/\sigma^2)
\end{aligned} \tag{A.7}$$

We consider two cases,

$$\epsilon/\sigma^2 \ll 1 \implies \left\langle \xi^2 \right\rangle \approx 1 - 2\rho \tag{A.8}$$

$$\epsilon/\sigma^2 \gg 1 \implies \left\langle \xi^2 \right\rangle \approx \frac{1}{1+\epsilon/\sigma^2} \approx \sigma^2/\epsilon \tag{A.9}$$

which allows for the approximation:

$$\left\langle \xi^2 \right\rangle \approx \frac{1}{\frac{1}{\sigma^2/\epsilon} + \frac{1}{1-2\rho}} \tag{A.10}$$

We may now substitute the approximation of $\left\langle \xi^2 \right\rangle$ in Eq. A.4,

$$\begin{aligned}
\overline{\Delta x^2} &\approx \frac{\epsilon}{\frac{1}{\sigma^2/\epsilon} + \frac{1}{1-2\rho} - 1} \\
&\approx \frac{\epsilon}{\frac{1}{\sigma^2/\epsilon} + 2\rho} \\
&= \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\epsilon/2\rho}}
\end{aligned} \tag{A.11}$$

Finally, substituting these values in the adaptive learning rate of Adadelta, we arrive at the approximation

$$\begin{aligned}
\eta_{\text{Adadelta}} &\approx \sqrt{\frac{\overline{\Delta x^2} + \epsilon}{\sigma^2 + \epsilon}} \approx \sqrt{\frac{\overline{\Delta x^2}}{\sigma^2}} \\
&= \sqrt{\frac{\frac{1}{\frac{1}{\sigma^2} + \frac{1}{\epsilon/2\rho}}}{\sigma^2}} \\
&= \frac{1}{\sqrt{1 + \frac{1}{\epsilon/2\rho/\sigma^2}}} \\
&= \frac{\sqrt{\epsilon/2\rho}}{\sqrt{\sigma^2 + \epsilon/2\rho}}
\end{aligned} \tag{A.12}$$

Comparing with the Rmsprop model, $\eta_{\text{Rmsprop}} = \eta^*/\sqrt{\sigma^2 + \epsilon^*}$, we can see that Adadelta is an approximation with a scaling meta-parameter $\eta^* = \sqrt{\epsilon/2\rho}$ and robustness meta-parameter $\epsilon^* = \epsilon/2\rho$.

We may also consider two cases, $\sigma^2 \gg \epsilon/2\rho$ and $\sigma^2 \ll \epsilon/2\rho$, which leads to an easily interpretable approximation. Adadelta performs variance normalization, with scaling $\eta^* = \sqrt{\epsilon/2\rho}$, and maximal learning rate 1:

$$\eta_{\text{Adadelta}} \approx \min\left(1, \frac{\sqrt{\epsilon/2\rho}}{\sigma}\right) \tag{A.13}$$

A.2 Numerical simulations

To test our analysis we simulate Adadelta for artificial data and compare it to our analytical results. We consider a stationary gradient with normal distribution, $g_t \sim N(0, \sigma^2)$, and we scan through a variety of values for the model parameters and gradient statistics, $\epsilon \in [10^{-8}, 10^{-2}]$, $\rho \in [10^{-3}, 10^{-1}]$ and $\sigma \in [10^{-3}, 10^{-1}]$. For each set of parameters, we estimate the average value of Adadelta, and plot it against the scaling factor $\frac{\sqrt{\epsilon/2\rho}}{\sigma}$ obtained in our approximation in Eq. A.12 (Fig. A.1). We can see that Adadelta has is well approximated as a function of the scaling factor, and that it is well fitted by the Rmsprop approximation.

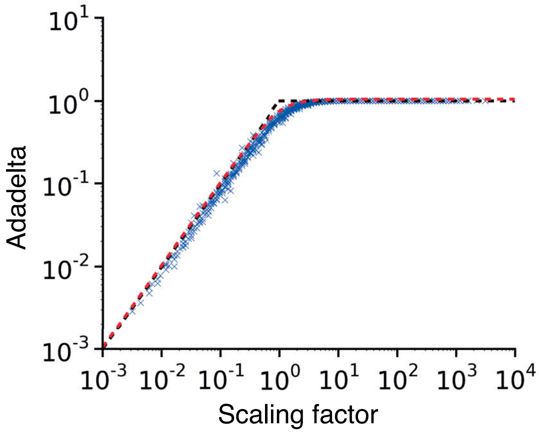

Figure A.1 – **Adadelta as an approximation of Rmsprop.** Average Adadelta values (blue marks) for different choices of meta-parameters and gradient statistics, plotted against the scaling factor obtained analytically, $\frac{\sqrt{\epsilon/2\rho}}{\sigma}$. The Adadelta values are well described as a function of the scaling factor, and is well approximated by the Rmsprop model (red dashed line), and the approximation in Eq. A.12 (black dashed line).

We conclude that Adadelta is an approximation of Rmsprop, implementing variance normalization, but not second-order optimization, as it was originally proposed. This analysis is also an illustrative application of the estimation theory we have developed, showing how adaptive models may be compared through their dependencies on gradient statistics.

Bibliography

L. Aitchison and P. E. Latham. Bayesian synaptic plasticity makes predictions about plasticity experiments in vivo. *arXiv:1410.1029 [q-bio]*, 2014.

R. A. Andersen, G. K. Essick, and R. M. Siegel. Encoding of spatial location by posterior parietal neurons. *Science*, 230(4724):456–458, 1985.

H. Barlow. Single units and sensation: a neuron doctrine for perceptual psychology. *Perception*, 1(4):371–394, 1972.

A. L. Barth and J. F. Poulet. Experimental evidence for sparse firing in the neocortex. *Trends in Neurosciences*, 35(6):345–355, 2012.

A. J. Bell and T. J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

N. Berardi, T. Pizzorusso, and L. Maffei. Critical periods during sensory development. *Current opinion in neurobiology*, 10(1):138–145, 2000.

P. Berkes, G. Orbán, M. Lengyel, and J. Fiser. Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. *Science*, 331(6013):83 –87, 2011.

E. L. Bienenstock, L. N. Cooper, and P. W. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience*, 2(1):32–48, 1982a.

E. L. Bienenstock, L. N. Cooper, and P. W. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience*, 2(1):32–48, 1982b.

B. S. Blais, N. Intrator, H. Shouval, and L. N. Cooper. Receptive Field Formation in Natural Scene Environments: Comparison of Single-Cell Learning Rules. *Neural Computation*, 10 (7):1797–1813, 1998.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

L. Buesing, J. Bill, B. Nessler, and W. Maass. Neural Dynamics as Sampling: A Model for Stochastic Computation in Recurrent Networks of Spiking Neurons. *PLoS Comput Biol*, 7 (11):e1002211, 2011.

T. Cai, J. Fan, and T. Jiang. Distributions of Angles in Random Packing on Spheres. *arXiv:1306.0256 [math, stat]*, 2013.

N. Caporale and Y. Dan. Spike Timing–Dependent Plasticity: A Hebbian Learning Rule. *Annual Review of Neuroscience*, 31(1):25–46, 2008.

J.-Y. Chen, P. Lonjers, C. Lee, M. Chistiakova, M. Volgushev, and M. Bazhenov. Heterosynaptic plasticity prevents runaway synaptic dynamics. *The Journal of Neuroscience*, 33(40):15915–15929, 2013.

A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surface of multilayer networks. *arXiv preprint arXiv:1412.0233*, 2014.

C. Clopath, L. Busing, E. Vasilaki, and W. Gerstner. Connectivity reflects coding: a model of voltage-based STDP with homeostasis. *Nature Neuroscience*, 13(3):344–352, 2010.

B. R. Conway and D. Y. Tsao. Color architecture in alert macaque cortex revealed by FMRI. *Cerebral Cortex*, 16(11):1604–1613, 2006.

L. N. Cooper and M. F. Bear. The BCM theory of synapse modification at 30: interaction of theory with experiment. *Nature Reviews Neuroscience*, 13(11):798–810, 2012.

L. N. Cooper, N. Intrator, B. S. Blais, and H. Z. Shouval. *Theory of Cortical Plasticity*. World Scientific Pub Co Inc, 2004.

M. C. Crair, D. C. Gillespie, and M. P. Stryker. The role of visual experience in the development of columns in cat visual cortex. *Science*, 279(5350):566–570, 1998.

Y. Dan, J. J. Atick, and R. C. Reid. Efficient coding of natural scenes in the lateral geniculate nucleus: Experimental test of a computational theory. *The Journal of Neuroscience*, 16(10): 3351–3362, 1996.

Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems 27*, pages 2933–2941. 2014.

Y. N. Dauphin, H. de Vries, J. Chung, and Y. Bengio. RMSProp and equilibrated adaptive learning rates for non-convex optimization. *arXiv:1502.04390 [cs]*, 2015.

H. A. David and H. N. Nagaraja. *Order statistics*. Wiley Online Library, 1970.

P. Dayan and L. F. Abbott. *Theoretical neuroscience*, volume 31. MIT press Cambridge, MA, 2001.

R. Desimone. Face-selective cells in the temporal cortex of monkeys. *Journal of cognitive neuroscience*, 3(1):1–8, 1991.

J. J. DiCarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

J.-R. Duhamel, F. Bremmer, S. BenHamed, and W. Graf. Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature*, 389(6653):845–848, 1997.

T. Elliott. Sparseness, antisparseness and anything in between: The operating point of a neuron determines its computational repertoire. *Neural computation*, pages 1–49, 2014.

D. Field. What is the goal of sensory coding? *Neural computation*, 6(4):559–601, 1994.

P. Foldiak. Forming sparse representations by local anti-hebbian learning. *Biological cybernetics*, 64(2):165–170, 1990.

D. J. Foster and M. A. Wilson. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440(7084):680–683, 2006.

F. Frassinetti, N. Bolognini, and E. Làdavas. Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research*, 147(3):332–343, 2002.

J. Freeman and E. P. Simoncelli. Metamers of the ventral stream. *Nat Neurosci*, 14(9):1195–1201, 2011.

J. H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82(397):249–66, 1987.

K. Friston. Beyond phrenology: what can neuroimaging tell us about distributed circuitry? *Annual review of neuroscience*, 25(1):221–250, 2002.

K. J. Friston, P. Fletcher, O. Josephs, A. Holmes, M. D. Rugg, and R. Turner. Event-related fMRI: characterizing differential responses. *Neuroimage*, 7(1):30–40, 1998.

R. C. Froemke and Y. Dan. Spike-timing-dependent synaptic modification induced by natural spike trains. *Nature*, 416(6879):433–438, 2002.

N. Frémaux, H. Sprekeler, and W. Gerstner. Functional Requirements for Reward-Modulated Spike-Timing-Dependent Plasticity. *The Journal of Neuroscience*, 30(40):13326–13337, 2010.

C. Fyfe and R. Baddeley. Non-linear data structure extraction using simple hebbian networks. *Biological Cybernetics*, 72(6):533–541, 1995.

W. Gerstner, R. Kempter, and J. L. Van Hemmen. A neuronal learning rule for sub-millisecond temporal coding. *Nature*, 383(6595):76–78, 1996.

W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski. *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge University Press, 2014.

M. Graupner and N. Brunel. Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location. *Proceedings of the National Academy of Sciences*, 109(10):3991–3996, 2012.

K. Grill-Spector, T. Kushnir, T. Hendler, S. Edelman, Y. Itzchak, R. Malach, and others. A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Human brain mapping*, 6(4):316–328, 1998.

C. G. Gross. Single neuron studies of inferior temporal cortex. *Neuropsychologia*, 46(3): 841–852, 2008.

C. Gulcehre and Y. Bengio. Adasecant: Robust adaptive secant method for stochastic gradient. *arXiv:1412.7419 [cs, stat]*, Dec 2014. arXiv: 1412.7419.

D. O. Hebb. *The organisation of behaviour: a neuropsychological theory*. Wiley, 1952.

T. K. Hensch. Critical period plasticity in local cortical circuits. *Nature Reviews Neuroscience*, 6 (11):877–888, 2005.

J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the theory of neural computation*, volume 1. Basic Books, 1991.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

G. E. Hinton, S. Osindero, and Y.-W. Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554, 2006.

J.-B. Hiriart-Urruty and C. Lemarechal. *Convex Analysis and Minimization Algorithms I: Fundamentals*. Springer Science & Business Media, 2013.

D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

S. A. Huettel, A. W. Song, and G. McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, 2004.

J. J. Hunt, P. Dayan, and G. J. Goodhill. Sparse coding can predict primary visual cortex receptive field changes induced by abnormal visual input. *PLoS Comput Biol*, 9(5):e1003005, 2013.

A. Hyvarinen and E. Oja. Independent component analysis by general nonlinear hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

A. Hyvarinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

A. Hyvarinen, J. Hurri, and P. O. Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.*, volume 39. Springer, 2009.

N. Intrator and L. N. Cooper. Objective Function Formulation of the BCM Theory of Visual Cortical Plasticity: Statistical Connections, Stability Conditions. *Neural Networks*, 5:3–17, 1992.

W. James. The principles ofpsychology (Vol. 1). *New York: Holt*, 1890.

E. R. Kandel, J. H. Schwartz, and T. M. Jessell. *Principles of neural science*. Appleton & Lange, 2000.

D. Kappel, S. Habenschuss, R. Legenstein, and W. Maass. Synaptic Sampling: A Bayesian Approach to Neural Network Plasticity and Rewiring. In *Advances in Neural Information Processing Systems*, pages 370–378, 2015.

M. Kaschube, M. Schnabel, S. Lowel, D. M. Coppola, L. E. White, and F. Wolf. Universality in the evolution of orientation columns in the visual cortex. *Science*, 330(6007):1113–1116, 2010.

P. D. King, J. Zylberberg, and M. R. DeWeese. Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of v1. *The Journal of Neuroscience*, 33(13):5475–5485, 2013.

D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, 2014.

H. Komatsu, Y. Ideura, S. Kaji, and S. Yamane. Color selectivity of neurons in the inferior temporal cortex of the awake macaque monkey. *Journal of Neuroscience*, 12(2):408–424, 1992.

J. Kominek and A. W. Black. *The CMU Arctic speech databases*. 2004.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

C. C. Law and L. N. Cooper. Formation of receptive fields in realistic visual environments according to the bienenstock, cooper, and munro (BCM) theory. *Proceedings of the National Academy of Sciences*, 91(16):7797–7801, 1994.

Q. V. Le. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8595–8598. IEEE, 2013.

N. Le Roux and Y. Bengio. Representational power of restricted Boltzmann machines and deep belief networks. *Neural computation*, 20(6):1631–1649, 2008.

Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10), 1995.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

H. Lee, C. Ekanadham, and A. Ng. Sparse deep belief net model for visual area v2. *Advances in neural information processing systems*, 20, 2007.

H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM, 2009.

S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.

J. C. Magee and E. P. Cook. Somatic EPSP amplitude is independent of synapse location in hippocampal pyramidal neurons. *Nature Neuroscience*, 3(9):895–903, 2000.

H. Markram, J. Lübke, M. Frotscher, and B. Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275(5297):213–215, 1997.

B. M. Mazoyer, N. Tzourio, V. Frak, A. Syrota, N. Murayama, O. Levrier, G. Salamon, S. Dehaene, L. Cohen, and J. Mehler. The cortical representation of speech. *Journal of Cognitive Neuroscience*, 5(4):467–479, 1993.

M. Merleau-Ponty. *Phénoménologie de la perception*. Gallimard, 1945.

K. D. Miller. A model for the development of simple cell receptive fields and the ordered arrangement of orientation columns through activity-dependent competition between on-and off-center inputs. *Journal of Neuroscience*, 14:409–409, 1994.

K. D. Miller, J. B. Keller, and M. P. Stryker. Ocular dominance column development: analysis and simulation. *Science*, 245(4918):605–615, 1989.

L. M. Miller, M. A. Escabi, H. L. Read, and C. E. Schreiner. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of neurophysiology*, 87(1):516–527, 2002.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

V. Mountcastle. An organizing principle for cerebral function: The unit model and the distributed system. 1978.

A. S. Nandy, T. O. Sharpee, J. H. Reynolds, and J. F. Mitchell. The fine structure of shape tuning in area V4. *Neuron*, 78(6):1102–1115, 2013.

R. Naud, N. Marcille, C. Clopath, and W. Gerstner. Firing patterns in the adaptive exponential integrate-and-fire model. *Biological cybernetics*, 99(4-5):335–347, 2008.

E. Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

E. Oja. Neural networks, principal components, and subspaces. *International journal of neural systems*, 1(01):61–68, 1989.

E. Oja, H. Ogawa, and J. Wangviwattana. Learning in nonlinear constrained hebbian networks. *Artificial Neural Networks*, 1991.

J. O'Keefe. A review of the hippocampal place cells. *Progress in neurobiology*, 13(4):419–439, 1979.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

B. A. Olshausen and D. J. Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487, 2004.

B. A. Olshausen and D. J. Field. How close are we to understanding V1? *Neural computation*, 17(8):1665–1699, 2005.

J. P. Pfister and W. Gerstner. Triplets of spikes in a model of spike timing-dependent plasticity. *The Journal of Neuroscience*, 26(38):9673–9682, 2006.

J.-P. Pfister, T. Toyoizumi, D. Barber, and W. Gerstner. Optimal spike-timing-dependent plasticity for precise action potential firing in supervised learning. *Neural computation*, 18(6): 1318–1348, 2006.

C. Pozzorini, R. Naud, S. Mensi, and W. Gerstner. Temporal whitening by power-law adaptation in neocortical neurons. *Nature Neuroscience*, 16(7):942–948, 2013.

C. J. Price. The anatomy of language: a review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, 1191(1):62–88, 2010.

A. A. Prinz, D. Bucher, and E. Marder. Similar network activity from disparate circuit parameters. *Nature Neuroscience*, 7(12):1345–1352, 2004.

R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

M. Rehn and F. T. Sommer. A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *Journal of Computational Neuroscience*, 22(2):135–146, 2007.

D. L. Ringach. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, 88(1):455–463, 2002.

M. Rivera-Alba, S. N. Vitaladevuni, Y. Mishchenko, Z. Lu, S.-y. Takemura, L. Scheffer, I. A. Meinertzhagen, D. B. Chklovskii, and G. G. de Polavieja. Wiring economy and volume exclusion determine neuronal placement in the Drosophila brain. *Current Biology*, 21(23): 2000–2005, 2011.

E. T. Rolls, A. Cowey, and V. Bruce. Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas [and discussion]. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 335(1273):11–21, 1992.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

C. J. Rozell, D. H. Johnson, R. G. Baraniuk, and B. A. Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural Computation*, 20(10):2526–2563, 2008.

D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, 1994.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5:3, 1988.

O. Sacks. *The man who mistook his wife for a hat: And other clinical tales*. Simon and Schuster, 1998.

C. Savin, P. Joshi, and J. Triesch. Independent component analysis in spiking neurons. *PLoS computational biology*, 6(4):e1000757, 2010.

A. Saxe, M. Bhand, R. Mudur, B. Suresh, and A. Y. Ng. Unsupervised learning models of primary cortical receptive fields and receptive field plasticity. *Advances in neural information processing systems*, pages 1971–1979, 2011.

A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

T. Schaul, S. Zhang, and Y. LeCun. No More Pesky Learning Rates. *arXiv:1206.1106 [cs, stat]*, 2012.

H. S. Seung. Learning in Spiking Neural Networks by Reinforcement of Stochastic Synaptic Transmission. *Neuron*, 40(6):1063–1073, 2003.

J. Sharma, A. Angelucci, and M. Sur. Induction of visual orientation modules in auditory cortex. *Nature*, 404(6780):841–847, 2000.

P. J. Sjostrom, G. G. Turrigiano, and S. B. Nelson. Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron*, 32(6):1149–1164, 2001.

P. J. Sjostrom, E. A. Rancz, A. Roth, and M. Hausser. Dendritic excitability and synaptic plasticity. *Physiol. Rev.*, 88(2):769–840, 2008.

E. C. Smith and M. S. Lewicki. Efficient auditory coding. *Nature*, 439(7079):978–982, 2006.

S. Song, K. D. Miller, and L. F. Abbott. Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, 3(9):919–926, 2000.

T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4, 2012.

G. Turrigiano. Too many cooks? intrinsic and synaptic homeostatic mechanisms in cortical circuit refinement. *Annual review of neuroscience*, 34:89–103, 2011.

I. V. Viskontas, R. Q. Quiroga, and I. Fried. Human medial temporal lobe neurons respond preferentially to personally relevant images. *Proceedings of the National Academy of Sciences*, 106(50):21329–21334, 2009.

T. P. Vogels, H. Sprekeler, F. Zenke, C. Clopath, and W. Gerstner. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science*, 334(6062): 1569–1573, 2011.

T. N. Wiesel and D. H. Hubel. Single-cell responses in striate cortex of kittens deprived of vision in one eye. *J Neurophysiol*, 26(6):1003–1017, 1963a.

T. N. Wiesel and D. H. Hubel. Effects of visual deprivation on morphology and physiology of cells in the cat's lateral geniculate body. *J Neurophysiol*, 26(978):6, 1963b.

D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

M. D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. *arXiv:1212.5701 [cs]*, 2012.

F. Zenke, G. Hennequin, and W. Gerstner. Synaptic plasticity in neural networks needs homeostasis with a fast rate detector. *PLoS computational biology*, 9(11):e1003330, 2013.

J. Zylberberg, J. T. Murphy, and M. R. DeWeese. A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of v1 simple cell receptive fields. *PLoS Comput Biol*, 7(10):e1002250, 2011.

Phone: +41 77 476 6300
E-mail: carlos.stein@epfl.ch
Lausanne, Switzerland

Carlos Stein Naves de Brito

Education

PhD, Computational Neuroscience, 2010 – present
EPFL – Lausanne, Switzerland
“Representation and learning in cortical networks”

M.S., Neuroscience, 2008 – 2009
University of São Paulo, Brazil
“Statistical models of connectivity between brain areas”

B.S., Computer Engineering, 2002 – 2007
Technological Institute of Aeronautics (ITA), Brazil

Research topics

Theory of cortical receptive field development
Biological correlates of deep neural networks
Variability and representation in cortical networks
Neural adaptation and efficient coding
Connectionist theories of subjective perception

Publications

Brito C.S.N., Gerstner W. (2015). “Nonlinear Hebbian learning as universal principle in receptive field development.” Submitted.

Brito C.S.N., Gerstner W. (2015). “Cortical synaptic plasticity as second-order invariant feature learning.” In preparation.

Baccalá L.A., Brito C.S.N., Takahashi D.Y., Sameshima K. (2013). “Unified Asymptotic Theory for All Partial Directed Coherence Forms.” Philosophical Transactions of the Royal Society A

Brito C.S.N., Baccala L.A., Takahashi D.Y., Sameshima K. (2010) “Asymptotic Behavior of Generalized Partial Directed Coherence.” Conf Proc IEEE Eng Med Biol Soc.

Talks and posters

Brito C.S.N., Gerstner W. (2015) ”Homeostatic synaptic depression explains receptive field development by implicit whitening”. Cosyne (poster)

Brito C.S.N., Gerstner W. (2014) ”A unifying theory of receptive field development”. Cosyne (talk)

Brito C.S.N. (2013) “Sparse coding with spiking neurons and Hebbian plasticity”. FACETS-ITN meeting (talk)

Hugues E., Brito C.S.N., Gerstner W., Romo R., Deco G. (2013). “A model of perceptual discrimination under sequential sensory evidence”. CNS (poster)

Brito C.S.N., Gerstner W. (2011) “General conditions for spiking neurons and plasticity rules to perform independent component analysis”. CNS (poster)

Teaching and supervision

Student project supervision, EPFL

Eszter Vertes, Master thesis project, 2014
"Top-down mechanisms in the visual cortex"

Gevorg Poghosyan, Master semester project, 2013
"Multilayer unsupervised learning with realistic networks"

Guillaume Riesen, Fullbright internship, 2012/13
"Development of V1 orientation maps with local plasticity rules"

Teaching assistant, EPFL, 2011-2014

Unsupervised and reinforcement learning in neural networks

Pattern classification and machine learning

Biological modeling of neural networks

Work experience

Intern, Google Inc., Mountain View, CA, 2006

Software development for large-scale data storage and retrieval in online advertisement systems

Intern, Aero-Spatial Institute (IAE), CTA, Brazil 2003

Development, parallelization and optimization of high-performance software for computational fluid dynamics

Courses and scholarships

Brains, Minds & Machines (CBMM) Summer Course, 2014.
Woods Hole, United States.

Advanced Course in Computational Neuroscience (ACCN)
2011. Bedlewo, Poland.

FACETS ITN - Marie-Curie PhD scholarship, 2010-2013.

DAAD Hochschulwinterkurs scholarship for German studies, 2005.
Freiburg, Germany.

Scientific outreach

Book chapter
Pessoa O., Brito CSN, "Neuroscience: the search for understanding brain and mind" to appear in "Science at the turn of the century" (2014), EDUFBa, Brazil

Seminar "How the brain represents the world". Art exposition "Arte della tavola", 2013. Lausanne, Switzerland.

Academic awards

Silver Medal on International Mathematical Olympiad (IMO), 2001

World Finalist on International Collegiate Programming Contest (ICPC), 2005,06

Gold on Brazilian Mathematical Olympiad, 2000,01,03,04

2nd prize on International Mathematics Competition, 2004,05

First prize on Brazilian Physics Olympiad, 2001

Languages

Portuguese, English, Spanish, French, German

Programming languages

Python, C++, Julia, Matlab